Coordinating sentence composition with error correction: A multilevel analysis

Mariëlle Leijten*°, Sven De Maeyer*+ & Luuk Van Waes*

* University of Antwerp, ° Flanders Research Foundation, + STATUA Centre for Statistics, Antwerp | Belgium

Abstract: Error analysis involves detecting and correcting discrepancies between the 'text produced so far' (TPSF) and the writer's 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. We hypothesize that error correction with speech recognition software differs from error correction with keyboard for two reasons. Speech produces auditory commands and, consequently, different error types.

The study reported on here measured the effects of (1) mode of presentation (auditory or visual-tactile), (2) error span, whether the error spans more or less than two characters, and (3) lexicality, whether the text error comprises an existing word. A multilevel analysis was conducted to take into account the hierarchical nature of these data. For each variable (interference reaction time, preparation time, production time, immediacy of error correction, and accuracy of error correction), multilevel regression models are presented. As such, we take into account possible disturbing person characteristics while testing the effect of the different conditions and error types at the sentence level.

The results show that writers delay error correction more often when the TPSF is read out aloud first. The auditory property of speech seems to free resources for the primary task of writing, i.e. text production. Moreover, the results show that large errors in the TPSF require more cognitive effort, and are solved with a higher accuracy than small errors. The latter also holds for the correction of small errors that result in non-existing words.

Keywords: Cognitive effort, dictation, error analysis, multilevel analysis, speech recognition, text produced so far (TPSF), text production, technology of writing.

journal of WRITING RESEARCH

Leijten, M., De Maeyer, S., & Van Waes, L. (2011). Coordinating sentence composition with error correction: A multilevel analysis. *Journal of Writing Research, 2* (3), 331-363.

1. Introduction

Revising and correcting errors in particular can be seen as demanding cognitive activities. Almost every article on writing includes a paraphrase on 'writing is cognitively demanding'. These highly demanding activities have been described in several models on working memory (Baddeley, 1986; Baddeley & Hitch, 1974; Kellogg, 1996; McCutchen, 1996). Writers need to juggle the constraints of the several subprocesses.

The writing subprocesses that are most relevant to this study are reading and editing. Editing signals *errors* in the output of planning, translating, programming, and executing. In this experiment we focus only on editing of the *text produced so far* (TPSF) that can be read on the computer screen (cf. difference between internal and external revisions, Lindgren & Sullivan, 2006). A recursive pattern of the 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.

The objective of the present study is to explain differences in revising behavior: what type of errors are immediately corrected, which other types 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 different experimental conditions. Based on previous research (Leijten, 2007a; Leijten, Van Waes, & Janssen, 2010), we hypothesize that some error types require 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.

1.1 Writing and writing modes

Writers use different writing modes to produce their texts: handwriting, keyboarding and speech recognition are the most widely used writing modes. Previous research has shown the influence of writing mode on the writing process (for a review, see Olive, Favart, Beauvais, & Beauvais, 2009; Van Waes & Schellens, 2003). We first describe some important characteristics of speech recognition. When writing with speech recognition the text is dictated in an auditory stream to the computer. The main strength of speech recognition lies in the combination of high speed text composition (via voice) and the appearance of text produced so far 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'. These kinds of (semantic) errors require extra monitoring and make it more difficult to benefit from the speed of composition (Honeycutt, 2003). Consequently, writers who use speech recognition for text production must revise intensively.

Writing and its subprocesses place a high demand on the storage and processing capacities of the working memory (Ransdell & Levy, 1999; Torrance & Galbraith, 2006). The logic of using speech recognition for writing texts 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 unclear, however, whether the execution characteristic of the writing mode is the most important characteristic writers benefit from[1]. For example, speech recognition generates only real words as errors because these items are part of the available lexicon while word processor errors can be typographical errors resulting in non-words.

Speech recognition is a hybrid writing mode that combines characteristics of classic dictating and word processing. Therefore, the text produced so far may play a different role in writing with speech recognition than it does in computer-based word processing. Previous studies show that classic 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 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. This specific characteristic of text production using speech recognition was transferred in the experimental design of the present study. However, before going into details about the study, we would like to frame this study in the context of error detection and correction.

---

[1] Most studies that 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.

1.2 Error detection and correction

During writing, errors may result from (linguistic or orthographic) rules, or they can be a discrepancy between the TPSF and one's mental representation of how the text should be. In other words, 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 are involved in detecting semantic and syntactic errors, but less so for typographic 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.

Revision during writing involves error analysis, comprising error detection, diagnosing and correction (Hayes, Flower, Schriver, Statman, & Carey, 1987). 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 et al., 1994; Larigauderie, Gaonac'h, & Lacroix, 1998; Piolat, Roussey, Olive, & Amada, 2004). Even more recently, Leijten et al. (Leijten, 2007a; Leijten & Van Waes, 2005, 2006) reported on various error correction strategies of professional writers who 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.

A related quantitative study (Leijten, Van Waes, & Ransdell, 2010) showed that writers not only differ 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 and mouse errors in the text immediately – possibly because there was no need to switch between writing modes to solve these repairs – while they were much more tolerant of

speech recognition errors. Other writers, however, were less tolerant of this type of error and often immediately solved almost all larger errors, typically 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 these 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 of a text and left unnoticed at the point of utterance.

The objective of the present study is to explain differences in revising behavior by providing an analysis of online text production that isolates the effects of writing mode from, accuracy, error span, and lexicality. The design includes the most frequently occurring error types found in a case study of professional writers. The error types were presented to college students who were asked to detect and correct errors in the TPSF and complete causal statements (no order was specified).

An analysis of a wide range of (online) measures as (a) interference reaction time, a measure of working memory resource allocation, (b) strategy, as measured by preparation time and tendency to correct errors immediately or after further text production, and (c) success as measured by accuracy to correct errors, provide information about the interaction with the TPSF in general and, more specifically, about mechanisms by which new writing technologies might constrain revision during writing. The following section explains the study in more detail and ends with an explanation of the hypotheses tested in this study.

2. Method

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 in a quasi-experimental setting.

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. 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.

2.1 Participants

Sixty students participated in this experiment. 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.

2.2 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 built according to a causal coherence relationship. All the materials were presented in Dutch. In 24 out of the 60 sentences used in the experiment, we varied four types of errors to construct the deficient TPSF based on error span, mode of writing and lexicality of the errors (Table 1).

Table 1. Classification of errors

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

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. 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 caused by misrecognition of the dictated text is:

(1) Spoken input 'I am writing a short text.'
(2) Incorrect output '**Eye** am writing a short text.'

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

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

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.

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

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). A more detailed description of the material can be found in Leijten, Ransdell & Van Waes (2010).

2.3 Design

The experiment employed a 2 (experimental condition speech versus non-speech) by 2 (two sets of sentences) within-subjects design (see Table 2). Two sets of sentences were constructed in which an equal number of errors was distributed in a comparable way. The type of errors was equally varied. The order in which these sets of sentences were presented to the participants was counterbalanced in the design of the experiment. In addition the sequence of how the sentences were offered was varied (non-speech and speech).

Table 2. Design of the experiment

Number of group	Order of experimental condition (speech, non-speech and sets 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

2.4 Procedure

The participants were assigned to groups of four to participate in the experiment. Before the participants started with the experiment, they were asked to put on a headset and position the button for the reaction test on the side of their non-dominant hand. 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 three parts:

1. Baseline Reaction Time
2. First part reading-writing test
3. Second part reading-writing test

Before a new part started, the participants systematically received a written and an oral introduction to the new task. 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). To log the linear development of the writing process during the completion task, Inputlog (Van Waes & Leijten, 2006; Van Waes, Leijten, & Van Weijen, 2009) was used to capture the keyboard & mouse input and calculate the pausing time afterwards.

The first 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, 2002; Olive, 2004). 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). When conducting a secondary task it is 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 second and the third 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.

2.5 Dependent variables

Five dependent variables were derived from the logging data of the experiment (via TPSF-program and Inputlog, more detailed information can be found in Leijten, Leijten, Ransdell, & Van Waes, 2010):

1. 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.
2. 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.
3. Production time: 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.
4. Delayed error correction: whether the cursor was initially either positioned within the TPSF clause that was presented as a writing prompt (immediate error correction), or after the clause (delayed error correction).
5. Accuracy: the accuracy represents the percentage of sentences with a (manipulated) error that was rewritten correctly.

2.6 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 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. However, we would also like to take a step back and consider whether there is any effect of the experimental speech condition on the (cognitive) interaction with TPSF's that are presented correctly, that is without any manipulated errors.

Hypothesis 1: mode of presentation effect (auditory versus visual-tactile) on the interaction with correct text produced so far (TPSF)

In the best possible use of speech recognition software for writing, all sentences should be produced correctly. The present experimental design takes advantage of the possibility of first isolating correct from incorrect sentences and therefore maximizes the possible positive impact of auditory input on memory load. The isolation of the correct sentences and the assumed positive effect of the auditory condition on the cognitive effort in the production task is the basis for the first hypothesis. We hypothesized that the addition of speech in the auditory condition could decrease the memory load of writers during the initial interaction with the correct TPSF.

This hypothesis is based on the presupposition that processing text via the auditory channel (speech recognition) requires fewer cognitive resources than via the visual-tactile channel (word processor), especially when no correction is involved (cf. Piolat, Olive, & Kellogg, 2005). Therefore, in this hypothesis we expect, for instance, that Interference Reaction Time, a measure of the time needed to turn attention to the secondary task of responding to an auditory probe while generating text, to be faster during speech presentation compared to the single visual-tactile presentation. Other measures of cognitive load (e.g. preparation time) should point in the same directions.

Hypothesis 2: effect of mode of presentation (auditory versus visual-tactile) on the interaction with incorrect text produced so far (TPSF)

Unfortunately, the ideal world as described above does not exist. At this time the state of the art of speech recognition is such that the errorless production of text is not possible (accuracy levels of expert users of speech recognition are up to 99% in an ideal situation, resulting in about one or two errors every five lines). Consequently, writers must develop compensation strategies for dealing with the errors in the TPSF (Leijten, 2007a, see chapters 2, 3 and 4; Leijten & Van Waes, 2005). Therefore, we formulate expectations for writing processes that contain errors and in which the already produced text does contain deficiencies.

The second hypothesis compares the effect of the TPSF that was either presented auditorially first (speech condition) or only visual-tactile (non-speech condition). Based on the results in previous research (Leijten, Van Waes, & Ransdell, 2010; Van Waes, Leijten, & Quinlan, 2010), we expect that errors that occur after the initial clause has been shown only visually, without speech, are more cognitive demanding than errors that occur after the context is offered both visually and auditory, with speech.

Hypothesis 3: effect of error type on the interaction with the text produced so far (TPSF)

In the final hypothesis we formulate our expectations related to the different kind of errors that appear in the deficient TPSF clauses. We distinguish two error types based on (a) error span (large vs. small errors), and (b) effect of lexicality (existing vs. non-existing words). Our expectations are mainly based on the (tendencies of the) results presented in previous research (Leijten, 2007b, see Table 3).

Table 3. Overview of hypothesis in relation to cognitive load

Hypotheses			
H3a	SR Large errors	>	SR Small errors
H3b	Existing words	>	Non-existing words

Hypothesis 3a: effect of error span (small vs. large errors)

This hypothesis 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.e. covering a character spread[2] of more than two characters), it may be easier to recognize the error, but at the same time, it may require more memory resources due to the time delay required for maintaining the difference in representation.

We expect that large error spans - more than two characters – lead to a higher cognitive load than small errors.

Hypothesis 3b: effect of lexicality (existing words versus non-existing words)

This hypothesis compares errors that involve lexicality (semantic level). Lexicality refers to whether the error is within a real word or a non-word. Errors within a real word can be meaning-based or surface-based while errors within non-words can only be surface-based. Because speech recognizers use a lexicon, they will generate, by definition, only existent, real words. In the normal course of events, non-existent words only occur in writing with keyboard & mouse and are caused by typing mistakes.

[2] 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 recoingition' (large error) on the other hand (cf. materials section).

The finding in Hacker et al. (1994) and Larigauderie et al. (1998) that spelling errors are easier to make than semantic errors, suggests that non-word errors should be easier to make than real word errors. Therefore, we predict that errors resulting in non-existent words - those that occur only in writing with keyboard and mouse - can be solved more efficiently than errors resulting in other existent words - that can occur in writing with speech recognition or, by chance, in keyboard & mouse.

3. Data analysis

The data from the TPSF experiment are analyzed from a hierarchical perspective. The application of multilevel analyses is mainly disseminated by Van den Bergh (Quené & Van den Bergh, 2004, 2008; Van den Bergh & Rijlaarsdam, 1996)[3]. We opted for this method because a unilevel approach leads to a possible loss of statistical power due to data aggregation on the participant's level resulting in one mean score per condition and per error type. These aggregated data do not always adequately treat differences between writers and between sentences when analyzing their behavior during the interaction with different error types in the TPSF (presented in two conditions). By aggregating we created data on how our respondents preferred to react to a TPSF of a certain kind, but we leveled out the possible individual differences between the sentences. Not taking into account this nuance can lead to an aggregation bias in the interpretation of the analyses. To avoid this aggregation bias (Bernstein, 1990) and fully take into account the within-writer and the within-sentence variance, multilevel analyses can be performed. Van den Bergh & Rijlaarsdam (1996) introduced multilevel analysis in writing research. They argue that multilevel models are often more powerful and that each observation can be nested within both individuals and sentences (trials characterized by error types and conditions). The result of this method is that each observation can be treated equally taking into account the differences between writers and sentence characteristics (Van den Bergh & Rijlaarsdam, 1996, p. 220). That the number of observations per person sometimes slightly differs[4], does not affect the power of the analyses. So, the main advantage of multilevel methods is that they account for the hierarchy within collected observations and the dependencies within a hierarchical structure (see also Goldstein, 1995).

---

[3] For a guide to multilevel analysis we would like to refer to a tutorial by Quené and Van den Bergh (2004).

[4] Although the number and type of sentences is strictly controlled in this experiment – as opposed to more ecologically valid observations of writing processes – a few sentences could not be added to the data set for technical reasons (e.g. when the starting time preceded the beep for the second task the reaction time was not taken into account because it did not intervene with the reading of the TPSF). The results for the variables related to these sentences were coded as missing values, which resulted in a data set with slightly deviating total numbers of scores.

For each variable (interference reaction time, preparation time, production time, immediacy of error correction, and accuracy of error correction), random cross-classified models are presented. Both sentences and writers are defined at the same level, and every response is nested within sentences and within writers. Each model consists of four parts: an estimated mean for that specific variable, a characterization of each writer and of each sentence (as a deviation of the mean), and a characterization of each TPSF response that can be seen as the interaction between sentence and writer (as a deviation from the mean of that writer and that sentence). This approach enabled us to analyze the effects of different conditions and error types on each dependent variable taking into account the variance between writers, the variance between sentences, and the interaction between writers and sentences related to the individual response to the TPSF-prompt. The random cross-classified model helps us to define whether, for instance, the participants are the most decisive factor in correcting various errors, or whether the error types have a (more) decisive influence on error correction.

We conducted the multilevel analysis in three steps (see Figure 1). In the first step we estimated the so-called 'zero model' to gain insight on the variance between participants, sentences and individual responses. These variances provide information about the distribution of the variance of participants as opposed to the variance of sentences and individual responses. By calculating the intra class correlations (ICC) we can evaluate the relative amount of variance that can be attributed to each of the elements in the model.

In the next step we estimated the 'net zero model' in which we have integrated the variables that characterize the participants. In this step we have analyzed whether specific characteristics of the participants are at play that need to be taken into account in further analyses. Based on these models we get insight in the unique variance between responses, after correction for higher level variables. It is this unique variance that we want to explain.

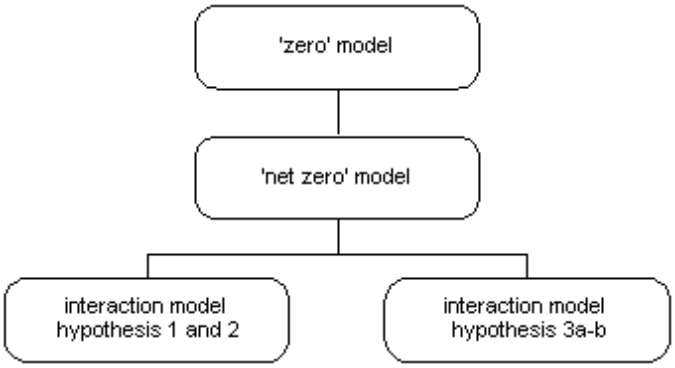

Figure 1. Flow of multilevel analysis.

In the final step we used interaction models (see section 3.2). The first type of interaction model is related to the experimental condition. Sentences that are offered only visually, versus sentences that are also presented via speech are compared for correct and incorrect sentences. The second type of interaction model compares the various error types that the writers are confronted with.

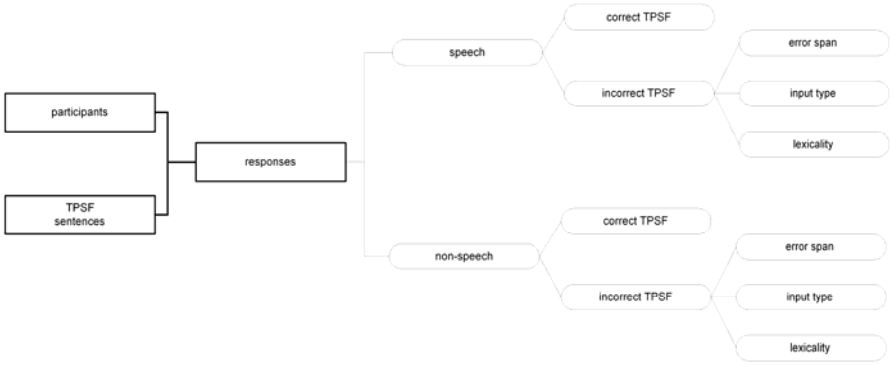

Figure 2. Hierarchical model describing the participant's level (level 2a) the sentence level (level 2b), and the response characteristics (level 1) used in the multilevel analysis.

In Figure 2 we describe the hierarchy between the different levels that were used in the multilevel analysis: the participant level (including participants' characteristics like, for instance, sex), the sentence level and the individual response level (nested within the previous two levels). The sentences/responses are characterized on the basis of the two experimental conditions, correctness and error types.

3.1 Characteristics of participants and sentences

In the analyses we estimate the effects of independent variables at participant and sentence level. The following explanatory variables were isolated for the participants:

1. Sex: male or female.
2. Groups: descriptive classification of error correction preference .
   We analyzed writers' preferences for correcting errors in the TPSF immediately versus delayed error correction. Therefore, we calculated the mean percentage of the cursor position in the incorrect sentences. We were able to describe four groups showing a distinct error correction preference. The median percentage of the cursor position is 88. We used this value as the breaking point to distinguish the so called 'immediate' from the 'delayed' group. Practically this immediate group has values ranging from 46 to 88%. The median of the immediate group is 65 and this again is used as the breaking value to split the immediate group into 'immediate' (46%-64%) and 'immediate medium' (65%-88%). The variance within those two groups is still relatively high. The delayed groups, on the other hand, are quite coherent.

The median and breaking point for this group is 98%. The participants prefer to complete almost every sentence before they switch to error correction. The 'delay medium' group ranges from 92% to 96%. The last group; the so called 'delay group', is almost unanimous in their preference: first they complete the text and then they correct the error.

3. Median baseline (this measure is only included for the variable reaction time): the median reaction time of the baseline test for each participant (cf. initial reaction time test) was added to the model as a residual to take the individual reaction differences into account.

Participants
- Sex (male or female)
- Groups immediate versus delayed
- Median baseline reaction time

Sentences
- Auditory condition (speech or non-speech)
- Correctness (correct or incorrect sentences)

Data focus
- Experimental condition (mode of presentation): correct (H1) and incorrect sentences (H2)
- Error type based on span (H3a)
- Error type based on lexicality (H3b)

Dependent variables describing cognitive effort
- Reaction Time
- Preparation Time
- Production Time
- Delayed correction
- Accuracy of correction (quality)

Figure 3. Variables used in the multilevel analyses.

The sentences might be influenced by the auditory condition (speech or non-speech) and by correctness. Figure 3 shows an overview of the variables that we have used in the analyses.

3.2 Models

Each model consists of a fixed part and a random part. The fixed part contains an estimated mean for that specific variable (β_0). The random part contains three terms: a unique characterization of each writer as a deviation of the mean ($u_{0(j0)}$), a characterization of each TPSF sentence ($v_{0(0k)}$), and a residual component that corresponds to the deviation of each individual response of a writer from its predicted value ($e_{i(jk)}$). In general the zero model, can be presented as follows:

$$Y_{i(jk)}=\beta_0 + [u_{0(j0)} + v_{0(0k)}+ e_{i(jk)}] \quad [1]$$

In this model $Y_{i(jk)}$ represents the observed score of observation *i* (*i*=1, 2,… I(*jk*)) of individual *j* (*j*=1, 2, …, J) and sentence *k* (*k*=1, 2, …, K). More specifically, in this study referring to a measure of cognitive effort: reaction time, preparation time, or production time for every response (i) nested within a participant (*j*) and a TPSF sentence (*k*).

Because in this study we used two types of variables to describe cognitive effort – continuous and binominal response variables (Rasbash, Steele, Browne, & Prosser, 2004) – we could not only use straightforward multilevel regression models as described above [Formula 1]. The binominal response variables (i.c. cursor position and accuracy) had to be fitted into another type of models, the so called logit multilevel regression model (Goldstein, 1995). It can be written as follows:

$$\text{logit}(\pi_{i(jk)}) = \beta_0 + [u_{0(j0)} + v_{0(0k)}] \qquad [2]$$

In this model $\pi_{i(jk)}$ is the probability that person *j* gets a score for sentence *k* of 1 on the dependent variable for response *i* . We used the logit link function because this enables us to translate estimates in odds ratios and calculate estimated probabilities based on these odds ratios.

In the next step we also added the variables that characterize the participants to the zero models (i.c. sex, counting span, median base line reaction time and preference for delayed correction). This resulted in so called 'net zero models'. This procedure enabled us to analyze which characteristics significantly influenced the value of the estimated means of the dependent variable. These characteristics are added in the interaction models to test our hypotheses. By doing this, we take into account possible disturbing person characteristics while testing our hypotheses.

To test the first two hypotheses we compared the mean values on our dependent variables for the mode of presentation and correctness of the TPSF (i.c. auditory condition [speech or non-speech] and correctness [correct or incorrect sentences]). To translate these assumptions in a model we combined the mode of presentation and correctness of the TPSF into four dummy variables ($D_{1i(jk)}$ - $D_{4i(jk)}$) identifying the characteristics per sentence. The interaction model (for the continuous variables) can be written as follows:

$$\begin{aligned} Y_{i(jk)} = {} & \beta_1 * D_{1i(jk)} + \beta_2 * D_{2i(jk)} + \beta_3 * D_{3i(jk)} + \beta_4 * D_{4i(jk)} + (\beta_5 * X_{1j} + \ldots + \beta_9 * X_{5j}) + \\ & [u_{10(j0)} + u_{20(j0)} + u_{30(j0)} + u_{40(j0)}] + [e_{1i(jk)} + e_{2i(jk)} + e_{3i(jk)} + e_{4i(jk)}] + [v_{0(0k)}] \end{aligned} \qquad [3]$$

$D_{1i(jk)}$ = non-speech - not correct
$D_{2i(jk)}$ = non-speech - correct
$D_{3i(jk)}$ = speech - not correct
$D_{4i(jk)}$ = speech - correct

In this model β_1 through β_4 are the estimates of the mean value on the dependent variable. For both the mode of presentation and correctness of the TPSF we estimate a residual at the person level ($u_{10(j0)}$ through $u_{40(j0)}$), the sentence level ($v_{0(0k)}$) and the response level ($e_{1i(jk)}$ through $e_{4i(jk)}$). Finally these models also contain variables at the respondent level (X_{1j} through X_{5j}), for which the effect sizes are estimated (β_5 through β_9).

To test the third hypothesis a comparable interaction model was built. In this model the different error types were added to the model to estimate the effect of error span and lexicality (see also Table 2). Because every error type was presented to the participants in both the speech and the non-speech condition, the model can be represented as follows:

$$\begin{aligned} Y_{i(jk)} = &\beta_1 * D_{1i(jk))} + \beta_2 * D_{2\,i(jk)} + \beta_3 * D_{3\,i(jk} + \beta_4 * D_{4i(jk)} + \beta_5 * D_{5i(jk)} + \beta_6 * D_{6\,i(jk)} + \\ &\beta_7 * D_{7\,i(jk)} + \beta_8 * D_{8\,i(jk)} + (\beta_9 * X_{1i(jk)} + \ldots + \beta_{13} * X_{5i(jk)}) + \\ &[u_{140(j0)} + u_{150(j0)} + u_{160(j0)} + u_{170(j0)}] + [e_{14i(jk)} + e_{15(jk)} + e_{16i(jk)} + e_{17i(jk)}] + [v_{0(0k)}] \end{aligned} \quad [4]$$

$D_{1i(jk)}$ and $D_{5i(jk)}$ = SR Large error resp. for non-speech (D_1) and speech (D_5)
$D_{2i(jk)}$ and $D_{6i(jk)}$ = SR Small error resp. for non-speech (D_2) and speech (D_6)
$D_{3i(jk)}$ and $D_{7i(jk)}$ = SR Small | Keyboard Small error resp. for non-speech (D_3) and speech (D_7)
$D_{4i(jk)}$ and $D_{8i(jk)}$ = Keyboard Small (non-existing) error resp. for non-speech (D_4) and speech (D_8)

We have chosen not to model separate variances for our 8 different conditions, but for four conditions. The reason therefore lies in a loss of precision of the estimates given the number of parameters to estimate in that case.

In the following section we present the results of the multilevel analyses used to test the three hypotheses put forward in this study.

4. Results

In a first step of the multilevel analysis procedure we explored the intra class correlation for each of the dependent variables we used to operationalize the memory load. The participants and the imbedded sentences differed for reaction time, preparation time and production time. The zero model estimates an intra class reaction time correlation of 37% at the participant level and about 60% at the sentence level. Therefore, it is advised to conduct multilevel analyses of the data. In the next step we added the various characteristics of the participants and the sentences to the zero model as to construct the netto zero model.

The participants' characteristics, sex, counting span and group do not have a significant effect on reaction time, but the median baseline that we have calculated for each participant on the basis of the initial reaction test does (estimated difference = 1.869, *SE* = 0.253). That is why we have taken into account the baseline reaction time in all analyses on reaction time. If we take the median baseline into account then the

ICC for participants slightly decreases from 37% to about 22% and for the sentence level it increases to about 74%. Table 4 shows the estimated means for the intercepts and the participants' characteristics that influenced the variables. The characteristics of the participants that differed significantly are taken into account in the further analyses.

The expected reaction time of an individual with an average median baseline is 397 *ms*. Preparation time is influenced negatively at participant level by the preference to correct errors immediately or to delay error correction (estimated mean = -360, *SE* = 54). Production time is influenced by the participants' sex. The female writers have significantly shorter production times than the male writers (estimated mean = -2732, *SE* =1038).

Table 4. Parameter estimates of intercept, participants' characteristics for reaction time, reparation time and production time

	Reaction time		**Preparation time**		**Production time**	
	Zero model	**Net zero model**	**Zero model**	**Net zero model**	**Zero model**	**Net zero model**
	***Est.* (*SE*)**	***Est.* (*SE*)**	***Est.* (*SE*)**	***Est.* (*SE*)**	***Est.* (*SE*)**	***Est.* (*SE*)**
Fixed Part						
Intercept	1338 (33)	397 (130)	1651 (90)	2168 (107)	17882 (777)	19339 (940)
Sex	-.-	-.-	-.-	-.-	-.-	-2732 (1038)
Counting span	-.-	-.-	-.-	-.-	-.-	-.-
Immediate/delayed error correction	-.-	-.-		-360 (54)	-.-	-.-
Median baseline RT	-.-	1.869 (0.253)	-.-	-.-	-.-	-.-
Random part						
Participant level Variance	49608 (9693)	24336 (5081)	332723 (68801)	171002 (39258)	17200918 (3252240)	15583367 (2953112)
Sentence level Variance	80739 (3115)	80741 (3115)	90814 (25816)	90584 (25731)	14825909 (3130905)	14826628 (3129767
Response level Variance	4376 (1694)	4339 (1679)	2050576 (55220)	2050555 (55219)	24098484 (648945)	24098510 (648946)
Loglikelihood	20371	19358	49961	49927	57213	57206
P		<.001		<.001		<.01

Est.= parameter estimate

SE = standard error

A calculation of the intra class correlations (ICC) indicates that the variance for production time at the participants' level and at the sentence level is quite comparable (resp. 30.6% and 26.4%). However, for preparation time and reaction time the ICCs are

less in balance. The time differences for preparation time at the individual level are about four times larger than at the sentence level (participant level 332723 (13.4%) vs. sentence level: 90814 (3.7%), resulting in a standard deviation of respectively 156 vs. 284). The differences for reaction time are in the opposite direction: the variance is larger at the sentence level than on the individual level: 59.9% (80739) of the variance is accounted for at the sentence level and 36.8% (49608) at the participant level. In other words, preparation time is more determined by the individual whereas the reaction time is more sentence determined.

Given that delayed error correction and accuracy are based on multilevel logit regressions, we conducted a Wald test (Rasbash et al., 2004) for the variance at participants' level to see if multilevel analyses are necessary. According to the Wald test on the variances at the participant level, delayed error correction needs to be analyzed with a multilevel model (χ^2=24.46, $p < .001$). The Chi-square for accuracy is 3.41, $p < .06$. Since this value is at the edge of the .05 significance level, we decided not to exclude this variable from the multilevel analyses. Table 5 shows the estimated means for the intercepts and the participants' characteristics that influenced the variables.

Table 5. Parameter estimates of intercept, participants' characteristics for delayed error correction and accuracy

	Delayed error correction		**Accuracy**	
	Zero model	**Net zero model**	**Zero model**	**Net zero model**
	Est. **(*SE*)**	***Est.*** **(*SE*)**	***Est.*** **(*SE*)**	***Est.*** **(*SE*)**
Fixed part				
Intercept	1.288 (0.169)	-.-	1.613 (0.279)	1.445 (0.127)
Delayed correction	-.-	-.-	-.-	0.408 (0.193)
Random part				
Participant level variance	1.655 (0.326)	-.-	0.172 (0.089)	0.070 (0.078)
Sentence level variance	0.064 (0.049)	-.-	1.601 (0.508)	1.601 (0.508)

Est. = parameter estimate

SE = standard error

Delayed error correction has an estimated mean of 1.288. None of the participants' characteristics has an influence on their preference to position the cursor in the TPSF or in the text completion part. The random part shows that the decision to position the cursor is largely related to individual characteristics of the participants (participant level: 1.655 (96.3%) vs. sentence level: 0.064 (3.7%)). The quality (accuracy) of the

texts on the other hand depends mainly on the sentence characteristics (participant level: 0.070 (9.7%) vs. sentence level: 1.601 (90.3%). In the netto model, however, we see that the accuracy scores are also influenced significantly by a particular participants' behavior, i.e. their preference to correct their errors immediately or to delay correction (delayed correction = 0.408, *SE* = 0.193).

4.1 Mode of error presentation effect on the interaction with correct TPSF

To describe the interaction between mode of presentation and the TPSF we have built 'interaction models'. The first formal interaction model [3] describes the effect of offering the correct or incorrect TPSF in an auditory and a visual condition (speech vs. non-speech). It comprises hypothesis 1 and 2. For the effect of the auditory condition on the interaction with corrected sentences we compared the estimates of parameters of $\beta_{1i(jk)}$ and $\beta_{2i(jk)}$.

In the first hypothesis we expected that the addition of speech in the auditory condition could have a positive effect on the memory load of writers. Table 6 shows the results for the speech versus non-speech condition on reaction time, preparation time and production time for correct sentences.

Table 6. Parameter estimates for the correct TPSF in the speech and non-speech condition[5]

	Speech		Non-speech			
	Est.	*SE*	*Est.*	*SE*	χ^2	sign. with 1df
Reaction time	465	73.2	464	72.3	< .01	1.000
Preparation time	1495	91.4	2455	122.7	72.13	< .001
Production time	15003	904.9	16087	16025.8	1.79	.181

The reaction time on the secondary task did not differ in the condition that the TPSF was also offered via speech. At first sight this is contrary to our predictions. We assumed that processing text via speech would require less cognitive resources and therefore result in faster reaction times. In both conditions the participants only needed about 465 milliseconds to respond (after correction of the median baseline RT). If however, we consider preparation time as another measure of cognitive effort, that is, the time it takes to start with the completion tasks, then we do see the difference that we expected in the first hypothesis. In the condition where writers first heard the TPSF read out loud before it was shown on the screen, they needed significantly less preparation time (speech = 1.495 seconds versus non-speech = 2.455 seconds preparation time (χ^2(1 df) = 72,13, $p < .001$). The difference in preparation time also affects the production time to complete the sentence. However it does not lead to a

[5] Because in this table we only report data that are related to 'correct' sentences and consequently no errors in the TPSF needed to be corrected, we do not report the values for cursor position and accuracy.

significant difference in the total production time. Text production is more or less of equal duration in the speech and the non-speech condition.

4.2 Mode of error presentation effect on the interaction with incorrect TPSF

The second hypothesis focuses on incorrect text. In general, we expect it to be easier to compare the mental representation of the TPSF with only the visual feedback on the screen, than to compare it with visual and auditory feedback. In short, to compare two things is easier than comparing three things. Above that, the auditory information of the TPSF probably leads to a focus on text production. In Table 7 the interaction with incorrect text is shown for the speech and the non-speech condition.

Table 7. Parameter estimates for the incorrect TPSF in the speech and non-speech condition

	Speech		Non-speech			
	Est.	***SE***	***Est.***	***SE***	χ^2	**sign.**
Reaction time	388	100.0	410	100.1	0.45	.502
Preparation time	1506	91.5	2787	152.1	78.52	< .001
Production time	19374	944.2	20563	1028.7	1.93	.165
Delayed error correction	2.72	0.39	0.67	0.21	22.10	< .001
Accuracy	1.41	0.30	1.41	0.29	0.00	1.00

If writers are confronted with errors in the TPSF, they still respond in the same way to the secondary task. In general, offering the text via speech has no influence on the reaction time when reading a TPSF clause in which an error occurs.

However when the sentences are read aloud, writers need less time to reflect on what their first writing action will be, in line with their behavior when dealing with correct sentences. The preparation time is significantly lower if the TPSF is also offered via speech (respectively 1.506 seconds compared to 2.787 seconds). When we compare the behavior of writers who prefer to correct the errors immediately or not, we see a significant difference in the use of preparation time. An additional analysis shows that the group of writers that prefers to correct the error in the text immediately in the non-speech condition needs three times more preparation time than the group of writers that prefer to delay error correction in the speech condition (preparation time non-speech immediate = 2.72 seconds versus preparation time speech delay = 0.87 seconds). These are the two most diverse patterns related to preparation time.

The production time is comparable in both conditions (about 20 seconds). The cursor position in incorrect sentences is significantly influenced by the spoken TPSF. The chance that writers who complete the sentence first is 31% higher in the speech

condition than in the TPSF (speech = 94% vs. non-speech = 66%)[6]. The odds of giving priority to text completion in speech are 15.24/1.95 = 7.81 times the odds compared to the situation in which the TPSF is not presented with speech[7].

The influence of the speech condition to either delay the error correction or not, seems to be different depending on the participants' general preference to postpone error correction or not. Figure 4 shows the preference for delayed error correction of the four writer groups (immediate, immediate medium, delayed medium, delayed). The graphical representation clearly shows the general tendency to delay the error correction when the TPSF is dictated first (speech-condition). Speech clearly reinforces the writers' preference to delay the error correction and to prioritize the completion of the sentence. However, Figure 4 also shows that the behavior within the non-speech condition is much more diverse than in the speech condition. In the former condition the group that – relatively spoken – is least eager to delay error correction in the speech condition (about 77%), delays three times fewer errors (about 24 %) in the non-speech condition. On the other hand there are participants who hardly solve any errors immediately, certainly not in the speech condition and hardly any in the non-speech condition (i.c. both the 'delayed' groups; max 10 %). The four groups are all characterized by a specific preference to delay error correction or not and their behavior differs significantly[8]. On the basis of these observations we can conclude that participants behave differently with respect to their preferred strategy to either delay errors or not, and that this behavior is significantly influenced by the occurrence of an auditory representation of the TPSF.

In general, speech does not influence the accuracy of the correction significantly, that is, the participants did not detect and correct more errors in the TPSF in either condition. However, an additional analysis shows that writers who prefer to delay error correction seem to be significantly more precise than writers who prefer to position their cursor first in the TPSF to start error correction immediately. The group that prefers to prioritize text production succeeds in 87% to correct the TPSF accurately and the group that prefers to revise first has an accuracy score of 81% (estimated difference: delay (delay + delay medium) = .436, *SE* = .171).

---

[6] The Chance were calculated on the basis of the reported beta scores as follows:
Chance[X] = 1/(1+(exp(-betascore(X))))*100

[7] The Oddsratios were calculated on the basis of the reported beta scores as follows:
Oddsratio[X] = exponent (betascore(X))

[8] Estimated differences:
'immediate medium'(67-89%)versus 'delayed medium'(90-96%): 1.62 and $\chi^2(1)=25.95$, $p < .001$;'delayed medium'(90-96%) versus 'delayed'(100%): 2.94 and $\chi^2(1)=10.58$, $p < .001$;
'immediate medium'(67-89%) versus 'delayed'(100%): 4.11 and $\chi^2(1)=13.78$, $p < .001$.

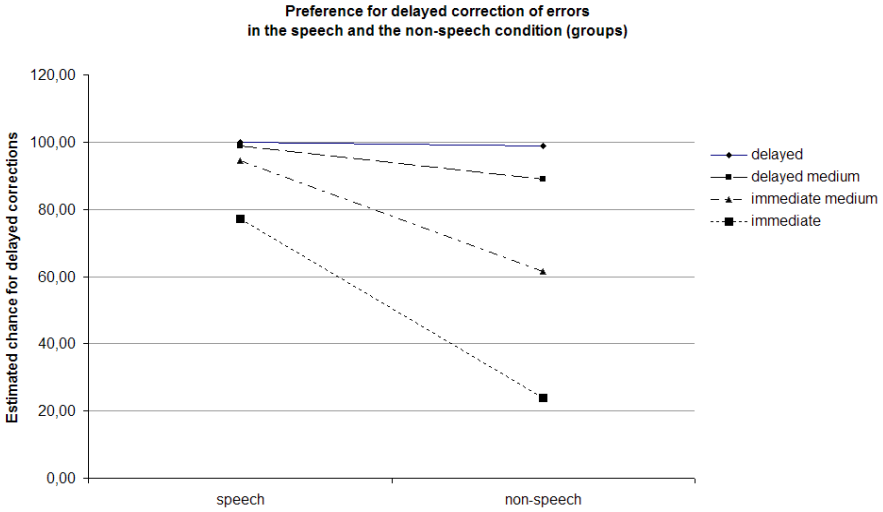

Figure 4. Preference for delayed correction of errors in the speech and non speech condition (groups).

The addition of offering the TPSF via an auditory channel affects the preparation time writers need to either complete the sentence or to start error correction, both in the ideal world with no errors in the TPSF and in texts that do contain deficiencies. Writers also have a strong tendency to continue text production if the TPSF is also presented in the speech mode. The next paragraphs describe the influence of error span and lexicality on error correction.

4.3 Effect of error span

The interaction model of the four error types that takes the interaction between mode of presentation and error types into account for the variable reaction time are described in Formula [4].

For the first hypothesis on error types we compared the effect of large speech recognition errors with small speech recognition errors on the memory load and the interaction behavior with the TPSF. For instance, the estimated means of the reaction time for large and small errors were compared in combination with the speech condition (SR Large Speech$_{ij}$ + SR Small Speech$_{ij}$). We expect that large error spans can be solved more efficiently than small errors. Table 8 shows the general behavior and the behavior in the speech and non-speech condition of writers in the interaction with large and small speech recognition based errors in the TPSF.

Table 8. Parameter estimates for Small and Large speech recognition errors

	SR Large Error		SR Small Error			
	Est.	***SE***	***Est.***	***SE***	χ^2	**sign.**
Reaction time	443	81.62	366	79.63	4.51	0.03
Speech°	394		344		0.88	0.35
Non-speech	493	83.95	388	80.29	6.11	0.01
Speech*error°°					1.57	0.21
Preparation time	2889	165.64	2159	91.02	18.90	0.00
Speech	1806		1612		0.77	0.38
Non-speech	4009	211.74	2713	108.03	34.10	0.00
Speech*error°°					12.75	0.00
Production time	26666	1365.57	19974	1231.83	14.00	0.00
Speech	25417		19778		9.31	0.00
Non-speech	27956	1419.02	20167	1261.04	34.10	0.00
Speech*error°°					12.75	0.00
Delayed error correction	2.47	0.21	1.60	0.23	0.47	0.49
Speech	3.06		2.72		0.37	0.54
Non-speech	0.69	0.24	0.98	0.26	0.69	0.41
Speech*error°°					1.39	0.24
Accuracy	2.08	0.57	0.89	0.49	2.53	0.11
Speech	1.90		0.93		1.56	0.21
Non-speech	2.29	0.62	0.85	0.51	3.32	0.07
Speech*error°°					1.00	0.32

° The values for speech are calculated based on the values of the estimate of the non-speech parameter.

°°The significance for the interaction terms are evaluated by comparing both differences between large and small errors.

The estimated reaction time is significantly longer with large errors than with small speech recognition errors (Large Error = 443 *ms* versus Small Error = 366 *ms*). The addition of speech causes a significant decrease of difference in reaction time (estimated difference: Large Error: -98.946; *SE* = 38.431 versus Small Errors: -44.528: *SE* = 20.101). Large errors distract more than small errors, but not when the visual TPSF is preceded by dictation.

In general, the error span does not interact with the mode of presentation, speech versus non-speech. Figure 5 shows the estimated means of reaction time for large and small errors in both conditions.

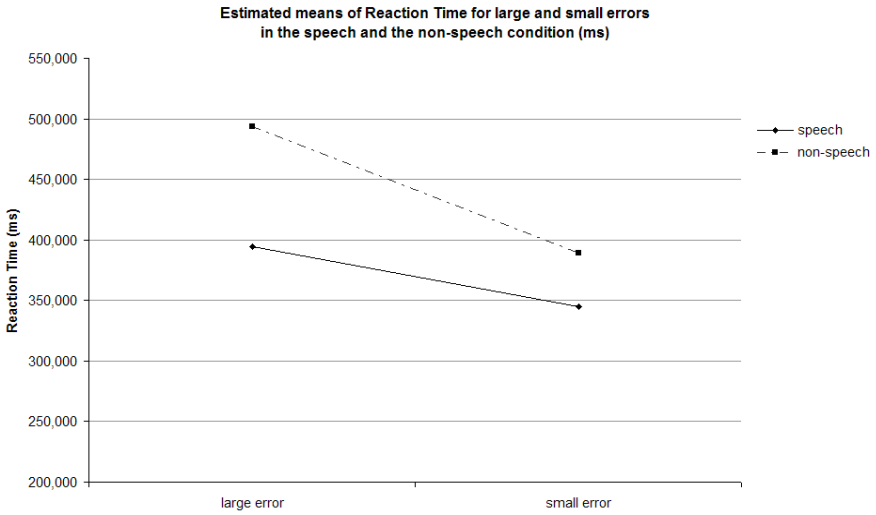

Figure 5. Estimated means of Reaction Time for large and small errors in the speech and the non-speech condition (ms).

As we can see in Figure 5, large errors result in significantly slower reaction times, either when the TPSF is also presented auditorially or not (speech vs. non-speech condition). The presence of large errors seems to distract the participants more intensively from their 'reading to produce' task and consequently increases the cognitive load for the writers in that stage of the writing process, especially in the non-speech condition. On top of that, the speech condition also lowers the cognitive effort significantly in these instances. The preference to correct errors immediately or to delay error correction does not influence the reaction time.

The preparation time is significantly longer when writers are confronted with a large error in the TPSF (Table 8). In the non-speech condition the preparation time to start correcting large errors (or completing a sentence on the basis of a TPSF with a large error) takes significantly longer than for sentences with small errors in the TPSF ($\chi^2(1) = 18.90$, $p < .001$).

The total production time (completion and correction) of sentences with a large error in the TPSF also takes significantly longer than for those with a small error (Large Errors = 27 seconds, Small Errors = 20 seconds). This may be partly explained by the larger amount of text production that needs to be performed by the writers who have to correct a larger error, but of course, also the level of distraction might take extra time to (mentally) reconstruct the correct TPSF. This result is most explicit in the non-speech condition ($\chi^2(1) = 34.10$, $p < .001$). If writers need to solve a large error in the TPSF, the prior dictation of the text seems to facilitate the writing process and significantly shortens the total text production time (estimated difference for speech condition with Large Errors: -2540; *SE* = 732). In other words, text production in these instances is

more fluent if the TPSF is also provided via the auditory channel. The interaction effect is significant. If we take a closer look at the small errors, then speech does not have any effect on the production time.

In general, the size of the error does not influence the preference of writers to start with error correction or to continue with text production. However, the speech condition does influence this preference: additional analyses on the difference between speech and non-speech show a significant estimated difference between the conditions (Large Errors = 2.47*;* *SE* = 0.21 and Small Errors=1.60; *SE* = 0,23). So, if the TPSF is offered via speech, the chance that writers prefer to delay error correction and complete the sentence first is higher.

Finally, in this experiment the size or span of the error does not seem to affect the probability that the error will be solved correctly or not. On average, the chance is about 85%. However, a more detailed analysis in which the preference to either delay the error correction or not is also taken into account, shows that writers who tend to delay more errors have a significant higher chance to correct more errors in the TPSF (estimated distance mean 0.473; *SE* = 0.158).

4.4 Effect of lexicality

In the final hypothesis on error types we focused on the effect of lexicality as a semantic characteristic of errors to compare the interaction with 'non-existing word' and 'existing-word' errors in the TPSF. We compared small errors that could either be caused by speech recognition software and keyboard based word processing with small errors that could only be caused by writing with keyboard. This latter type of error resulted in non-existent words, while the former error type consisted only of existing words. We expected that the non-existing words would be corrected more efficiently. Table 9 shows the parameter estimates for both error types.

The difference between existing words and non-existing words does not seem to causes any difference. The cognitive effort it takes to solve these errors does not seem to vary significantly which is comparable to the result of the previous hypothesis. Lastly, we did not find any interaction effects.

Table 9. Parameter estimates for small speech recognition/keyboard errors versus small keyboard errors

	Existing words SR small | Keyboard Small		Non-existing words Keyboard Small			
	Est.	*SE*	*Est.*	*SE*	χ^2	sign.
Reaction time	425	80.88	413	80.25	0.06	0.81
Speech°	455		412		0.65	0.42
Non-speech	396	82.69	415	81.27	0.14	0.71
Speech*error°°					2.22	0.14
Preparation time	2234	87.31	2202	90.19	0.12	0.73
Speech	1737		1780		0.13	0.72
Non-speech	2741	102.86	2636	104.39	0.76	0.38
Speech*error°°					1.02	0.31
Production time	19654	568.19	19548	1221.31	0.01	0.99
Speech	19789		19130		0.00	1.00
Non-speech	20207	1223.14	19969	1244.74	0.02	0.89
Speech*error°°					0.19	0.66
Delayed error correction	1.37	0.21	1.17	0.20	1.48	0.22
Speech	2.77		2.80		0.13	0.72
Non-speech	0.65	0.24	0.39	0.24	0.56	0.45
Speech*error°°					0.02	0.89
Accuracy	1.11	0.49	2.03	0.50	1.72	0.19
Speech	1.12		2.09		1.75	0.19
Non-speech	1.11	0.51	1.96	0.33	1.38	0.24
Speech*error°°					0.07	0.79

° The values for speech are calculated based on the values of the estimate of the non-speech parameter.

°°The significance for the interaction terms are evaluated by comparing both differences between large and small errors.

5. Conclusion and discussion

The present research isolated the effects of the writing mode that presented the TPSF (auditory vs. visual-tactile) from error span (large vs. small errors) and lexicality (existing vs. non-existing words). The data were explored via mixed-effects multilevel analysis. Although study's central focus was on error correction strategies, we have deliberately taken a step backwards and also considered the effect of the auditory channel in correct sentences. This enabled us to focus on the effect of the presentation mode.

The isolated correct sentences did not show a significant positive effect on reaction time in the speech and non-speech condition when the TPSF was also presented auditorially. However, the preparation time writers needed to complete the sentences

did drop significantly. Consequently, the decrease of cognitive effort is not fully confirmed (H1). However, we have to take into account that preparation time is more determined by the characteristics of the individual, whereas reaction time is mainly sentence determined. Furthermore, for preparation time the effect sizes are strong (1.76 for persons and 2.45 for sentences). The effect size of the reaction time is also quite strong for the participants (0.49), but small to medium for sentences (0.27). Therefore, we might conclude that the person characteristics that we have taken into account so far proved to be rather weak predictors. In order to better distinguish between writers it might therefore be advisable to address more specific person characteristics (for instance, related to the cognitive capacity of the writers, cf. infra) in combination with a variation in the task complexity.

Another explanation could be that the nature of the experimental task might have influenced the participants' behavior in dealing with correct sentences. Because they knew that an error *could* occur in the TPSF, we might have created a situation in which participants might have been looking more carefully to the correct sentences than they would in a normal situation, because they wanted to be sure that they had not overlooked the implemented error. In other words, it is possible that we have invoked an artificial evaluative, seeking reading behavior in the correct condition. So the attempt to create a bias towards correct TPSF sentences in the experimental design might not have worked out completely. Therefore, we suggest a follow-up study in which only correct sentences are provided. That would exclude possible noise in the experimental setup and provide solid evidence on the capabilities of speech to free resources.

The experimental condition in which the TPSF was either preceded by speech or not , influences the writer's strategies during error analysis. When isolating the incorrect sentences, we notice that writers more often delay error correction in the speech condition, and start writing sooner (H2). When speech proceeds the TPSF, writers can overtly compare the TPSF when it appears on screen with the speech. However, in the non-speech condition, only an internal, covert conflict is possible. Sometimes, speech 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 slowing down starting time. In other words, the auditory channel does cause a significant focus on fluent text production.

Compared to the mode of error presentation, error span has a more consistent effect on strategy choice. Writing without speech, and with large errors, leads to the highest cognitive effort in error analysis (H3a). Large errors lead to slower preparation time, longer production times, and slower interference reaction times, indicating that they consume more working memory resources. A positive effect of speech can definitely be found in the comparison between large and small errors. In general large errors distract more than small errors, but not when the TPSF is also offered via the auditory channel. Writers also need less preparation time when speech is present. It even does not make

a difference if the error is large or small. If we describe fluency as a measure to continue text production, than the fluency is significantly higher in the speech than in the non-speech condition: writers more often prefer to continue text production in a fluent way when the TPSF is dictated first.
The error types used in this experiment can be divided into three categories: large errors, small errors that result in existing words and small errors that result in non-existing words. The last two error types(existing words and non-existing words) do not seem to cause any difference (H3b). However, the large errors compared to the smaller errors distract writers in a way that it takes them longer to continue with text production or correct the error. So, large errors are cognitively demanding, but are accurately solved.

In general, we expected it to be easier to compare the mental representation of the TPSF with only the visual feedback on the screen, than to compare it with visual and auditory feedback. In short, to compare two things is easier than comparing three things. The results provide evidence that writers conduct three tasks differently than two tasks. It seems that writers opt not to conduct three activities. When speech is present the preparation time is shorter and writers generally prefer text completion, which suggests that they use the auditory information merely to continue text production. A similar experiment with eyetracking confirms this assumption (Van Waes, Leijten, & Quinlan, 2010). Writers can either continue text production based on the auditory information or they can use the visual information as a trigger to continue text production. Error correction is not a priority in this situation; the TPSF is just a vague visual stimulus to continue text production.

The preference to correct errors immediately or to delay error correction has no influence on the reaction time. An explanation might be that writers opt for a correction strategy that is most related to their working memory capacity. If writers expect delaying error correction might cause an extra burden, they more often opt to correct the error first, and vice versa. When writers are asked about the rationale behind the strategy of delaying the correction of an error, they state that the subprocess of production is sometimes more important and that text completion is performed before error correction because they are afraid to loose the 'gist' of their formulation. Above all, the cognitive planning process related to the content development was almost removed in this experiment. In a follow up study, it might be helpful to integrate a more complex planning component by providing the context not as a full sentence, but, for instance, only as keywords that need to be integrated in a self composed sentence clause. Moreover, it is advisable to take into account the effect sizes reported for the variable reaction time. Based on the results it seems to be advisable to incorporate an even larger set of sentences in order to create greater statistical power.

In this experiment we described the cognitive processes of writers via the variables' reaction time, preparation time, production time, delayed error correction and accuracy. A well-known measure of cognitive load during writing processes is reaction time. In using this measure it is important to choose the exact moment the secondary task is required. Since the sentences that needed to be read were rather short, the variation in offering the secondary task was limited to a small span. As not to bias the writers, we varied the timing of the second task in the correct filler sentences on a broader scale. However, the reaction time was not as decisive as we assumed. Only in the most extreme writing situation - large speech recognition errors – did it provide more insight in the cognitive effort. In a more neutral writing situation this measure was inconclusive. The supplementary measure of preparation time seemed to be more informative for this experimental setup. The time participants needed to decide what their next writing action would be differed as well for the mode of presentation and for the comparison between large and small errors. Since the task included the instruction 'speed on task' we see this measure as highly informative for the cognitive effort it takes to continue the writing process (whether this is production or correction). The final time measure is production time. Shorter production time is related to easier writing processes. Writers that are more fluent in text production produce longer text in the same amount of time as less fluent writers. Therefore, production time can also be taken as a measure of cognitive effort: the cognitive effort it takes to produce a text as fast and accurately as possible. The cursor position that writers choose can be described as a strategy choice to continue text production or to revise first. Again this can be seen as a cognitive effort measure. If speech is present, writers prefer to continue text production. In a follow up study in which writers were forced to correct errors first, the production time was significantly lower (Quinlan, Loncke, Leijten, & Van Waes, 2009). In this study, reaction time as 'the measurement' of cognitive effort *an sich*, would not have provided as much information as the combination of the time, strategy and accuracy measures together. In our opinion the combinations of various measurements are needed to accurately describe the cognitive processes in writing (Leijten, 2007a; Van Waes, Leijten, & Quinlan, 2010).

Acknowledgements

We especially would like to thank Isabelle De Ridder for co-designing and coordinating the experimental sessions, and data collecting. The fruitful discussions with Sarah Ransdell laid the foundation for this research project. Bart Van de Velde did a great job in programming the experiment.

The project was funded as a BOF/NRI (New Research Initiatives) project by the University of Antwerp. The logging tool Inputlog is available online: www.inputlog.net.

References

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.

Bernstein, L. (1990). *Developing an adequately specified model of state level student achievement with multilevel data*. Paper presented at the American Educational Association

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.

Goldstein, H. (1995). *Multilevel statistical analysis*. London: Edward Arnold.

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). 'Seeing it on the screen isn't really seeing it': Computer writers' reading problems. In G. E. Hawisher & C. L. Selfe (Eds.), *Critical perspectives on computers* (pp. 16-29). New York: Teachers College Press.

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.

Hayes, J. R., Flower, L., Schriver, K., Statman, J., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Reading, writing, and language possessing* (Vol. 2, pp. 176-240). Cambridge: Cambridge University Press.

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

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. (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. (2007a). How do writers adapt to speech recognition software? The influence of learning styles on writing processes in speech technology environments. In M. Torrance, L. Van Waes & D. Galbraith (Eds.), *Writing and Cognition: Research and Applications* (Vol. 20, pp. 279-292). Oxford: Elsevier.

Leijten, M. (2007b). *Writing and speech recognition: observing error correction strategies of professional writers* (Vol. 160). Utrecht: LOT.

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.

Leijten, M., Van Waes, L., & Janssen, D. (2010). Error correction strategies of professional speech recognition users: three profiles. *Computers in Human Behaviour, 26,* 964–975.

Leijten, M., Van Waes, L., & Ransdell, S. (2010). Correcting Text Production Errors: Isolating the Effects of Writing Mode From Error Span, Input Mode, and Lexicality. *Written communication, 27*(2), 189-227.

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. (2002). 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. (2006). Analyzing on-line revision. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer Key-Stroke 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., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisation. *Learning and Instruction, 19*(4), 299-308.

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

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*(3), 291-312.

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.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413-425.

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.

Quinlan, T., Loncke, M., Leijten, M., & Van Waes, L. (2009). *Writers' shift between error correction and sentence composing: Competing and the executive function*. Antwerp: University of Antwerp.

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., & 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.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). A user's guide to MLwiN Version 2.0. Retrieved February 6 2007, 2007

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.

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.

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., & Leijten, M. (2006). Logging writing processes with Inputlog. In L. Van Waes, M. Leijten & C. Neuwirth (Eds.), *Writing and Digital Media* (Vol. 17, pp. 158-166). Oxford: Elsevier.

Van Waes, L., Leijten, M., & Quinlan, T. (2010). Reading during sentence composing and error correction: A multilevel analysis of the influences of task complexity. *Reading and Writing: An Interdisciplinary Journal*, 23(7), 803-834.

Van Waes, L., Leijten, M., & Van Weijen, D. (2009). Keystroke logging in writing research: Observing writing processes with Inputlog. *GFL-German as a foreign language, 2*(3), 41-64.

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.