The effectiveness of comprehensive corrective feedback in second language writing
van Beuningen, C.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 4

Evidence on the effectiveness of comprehensive error correction in second language writing

Large-scale study

4.1 Abstract

The current study investigated the effect of direct and indirect comprehensive corrective feedback (CF) on second language (L2) learners' written accuracy (N = 268). The study set out to explore the value of CF as a revising tool as well its ability to constitute long-term accuracy development. In addition we tested Truscott’s (2001; 2007) claim that correction may have value for non-grammatical errors, but not for errors in grammar, and explored the structural complexity and lexical diversity of learners’ writing to see if students are inclined to avoid more complex constructions due to error correction (Truscott, 2007). Results showed that both direct and indirect CF led to improved accuracy, not only during revision but also in new pieces of writing (i.e. texts written during post-test and delayed post-test sessions). Furthermore, a separate analysis of grammatical and non-grammatical error types revealed that whereas only direct CF initiated grammatical accuracy gains, pupils’ non-grammatical accuracy benefited most from indirect CF. Moreover, CF did not result in simplified writing. These findings suggest that comprehensive CF is a useful instrument that teachers can use to help L2 learners improve their written accuracy over time.

1 A slightly adapted version of this chapter has been accepted for publication as: Van Beuningen, C.G., De Jong, N. H., & Kuiken, F. (accepted for publication). Evidence on the effectiveness of comprehensive error correction in second language writing. Language Learning.
4.2 Introduction

Error correction or CF is probably the most widely used feedback form in present day L2 classrooms. Its usefulness however, has been a fiercely debated topic ever since Truscott’s 1996 article in which he claimed error correction to be necessarily ineffective and potentially harmful. In the decade that followed, he has repeatedly presented objections with respect to the use of CF in L2 writing classes (Truscott 1996; 1999; 2004; 2007; 2009).

Truscott’s (1996) statement that CF is ineffective relies on both practical and theoretical arguments. His practical doubts pertain to teachers’ capacities in providing adequate and consistent feedback, and learners’ ability and willingness to use the feedback effectively. His argument against the effectiveness of error correction furthermore rests on the claim that CF overlooks important insights from second language acquisition (SLA) theories. First, Truscott denied CF to have any potential merit because of the mismatch between error correction as a simple transfer of information, and the gradual and complex nature of interlanguage development. He furthermore argued that since morphological, syntactic, and lexical knowledge are not all acquired in the same way, no single form of CF can be expected to be effective in all of these domains. In fact, Truscott denied CF to have any potential value for the development of grammatical competence, because – in his view – CF will be unable to affect the rules underlying syntactic and morphological errors. He suggested that CF could only be beneficial for errors that “are relatively simple and can be treated as discrete items” such as spelling errors, but never for grammatical features which are “integral parts of a complex system” (Truscott, 2007, p. 258). Third, and related to the previous argument, Truscott’s view that CF fails to affect the developing interlanguage system, led him to suggest that CF could at best help in developing explicit declarative knowledge, but never the more important implicit procedural knowledge (e.g. DeKeyser, 2003; Ellis, 2004). He deduced that CF could only lead to ‘pseudo-learning’; it might improve learners’ self-editing and revision skills, but could not result in accuracy development. The fourth and final theoretical objection raised by Truscott concerns the so-called ‘readiness problem’ (Pienemann, 1998). He explained that CF can only be expected to have any value when it is aligned to a learner’s current level of L2 development, but that insights into developmental sequences to date are too limited to be useful for teaching practice. From this Truscott concluded that L2 learners are not likely to benefit from CF.

Besides arguing that error correction is ineffective (for the reasons listed above), Truscott (1996; 2004; 2007) stated that correcting students’ writing could even be counterproductive. One of his arguments was that teachers run the risk of making their
students avoid more complex structures when they emphasize learners’ errors by providing CF. Truscott reasoned that it is the immediate goal of error correction to make learners aware of the errors they committed, and that this awareness creates a motivation for students to avoid the corrected constructions in future writing (Truscott, 2007). Secondly, Truscott (1996; 2004) claimed CF to be a waste of time, and suggested that the energy spent on dealing with corrections – both by teachers and students – could be allocated more efficiently to alternative activities, such as additional writing practice.

Based on the above, Truscott (1996) summoned the abandonment of CF from L2 writing classes, until its usefulness had been proven by empirical research (Truscott, 1999; 2004; 2007). Although Ferris (1999; 2002; 2004) made a stand for the use of written CF and argued that Truscott’s conclusions were premature, she agreed that evidence from well designed studies was necessary before any firm conclusions could be drawn about the (in)effectiveness of error correction. This call has resulted in an ever expanding body of studies exploring the effects of CF on L2 learners' writing.

Framed within a cognitive perspective on L2 acquisition, the present study reports on a tightly controlled classroom-based quasi-experiment, which included pre-test, treatment, post-test, and delayed post-test sessions, and compared the effects of direct and indirect comprehensive CF on the written accuracy exhibited by 268 secondary school L2 learners during revision as well as in new writing, one and four weeks after the treatment. The study also examined the differential value of CF for grammatical and non-grammatical error types, as well as any putative complexity avoidance effects of the CF treatments.

4.3 Literature review

4.3.1 Research into the effectiveness of written CF

The value of CF for revising and language learning

Earlier empirical work on the effects of CF on L2 learners writing could be categorized into two strands. While the first set of studies focused on the role of CF during the revision process (e.g. Ashwell, 2000; Fathman & Whalley, 1990; Ferris, 1997; Ferris & Roberts, 2001), the second group of investigations set out to answer the question if correction yields a learning effect (e.g. Chandler, 2003; Kepner, 1991; Polio, Fleck, & Leder, 1998; Semke, 1984; Sheppard, 1992).
Students in the revision studies demonstrated that they were able to improve the accuracy of a particular piece of writing, based on the feedback provided (e.g. Ashwell, 2000; Fathman & Whalley, 1990; Ferris, 1997; Ferris & Roberts, 2001). These findings thus showed that CF is a useful editing tool. However, Truscott and Hsu (2008) rightly argued that these results do not constitute evidence of learning. Whereas measuring a learning effect “necessarily involves a comparison between two independently written works” (Truscott & Hsu, 2008, p. 293), the revision studies mentioned above, only compared two versions of the same manuscript. It thus remains unclear if students’ success in using the feedback during revision will subsequently lead to acquisition of the corrected forms.

The more interesting body of research consists of studies that investigated the effect of CF on new pieces of writing. The first studies that did opt to provide insights into the learning potential of CF, however, produced mixed results (e.g. Chandler, 2003; Kepner, 1991; Polio et al., 1998; Semke, 1984; Sheppard, 1992). Whereas Chandler (2003) concluded that CF is an effective means of improving the accuracy of students’ writing over time, Kepner (1991), Semke (1984), Polio et al. (1998), and Sheppard (1992) failed to find an effect of error correction. All of these studies, however, suffered from serious shortcomings in terms of their design, execution, and/or analyses (see Chapter 3 or Guenette, 2007 for a review of these methodological issues). These investigations thus did not allow for any firm conclusions on the role of CF in L2 accuracy development, and led both opponents (e.g. Truscott, 1996) and advocates (e.g. Ferris, 1999) of written error correction to call for more, well designed CF studies.

The language learning potential of focused and unfocused CF
The above mentioned appeal has resulted in a growing body of tightly controlled investigations, exploring the long-term effects of CF on L2 writing, by comparing learners’ accuracy performance on pre-tests and (delayed) post-tests (see Chapter 3; see also Bitchener, 2008; Bitchener & Knoch, 2008; Bitchener & Knoch, 2009; Bitchener & Knoch, 2010a; Bitchener & Knoch, 2010b; Ellis, Sheen, Murakami, & Takashima, 2008; Sheen, 2007; Sheen, 2010b; Truscott & Hsu, 2008). When considering the type of feedback under investigation, these studies fall into three groups: those evaluating the effectiveness of focused CF, those examining the effects of unfocused CF, and those comparing the efficacy of focused and unfocused CF approaches. The focused-unfocused dichotomy refers to the comprehensiveness of CF methodologies. Whereas the unfocused or comprehensive approach involves correction of all errors in students’ texts, independent of
their error category, focused or selective CF targets a (number of) specific linguistic feature(s) only. Errors outside the focus domain are left uncorrected.

Research into the effectiveness of focused CF

As opposed to the earlier CF work discussed in the previous section, and following the methodology of oral feedback research (e.g. Lyster, 2004; Ellis, Loewen, & Erlam, 2006), the majority of recent written CF studies explored the effects of focused correction on learners’ accuracy development (e.g. Bitchener, 2008; Bitchener & Knoch, 2008; Bitchener & Knoch, 2009; Bitchener & Knoch, 2010a; Bitchener & Knoch, 2010b; Ellis et al., 2008; Sheen, 2007; Sheen, 2010b). In these studies, CF targets one persistently problematic error type only (e.g. errors in the use of English articles). The rationale behind this approach is that learners might be more likely to notice and understand corrections when just one feature is targeted (Ellis et al., 2008). Moreover, it was hypothesized that focused CF should be expected to be more effective than unfocused CF, because learners have a limited processing capacity (Bitchener, 2008; Sheen, 2007). The proposition is that, asking students to deal with CF which targets a broad range of linguistic features at the same time, may produce a cognitive overload, and prohibit feedback processing. Studies investigating the effectiveness of a focused approach to error correction, all reported robust and durable positive effects of CF on learners’ accuracy development (see Xu, 2009, however, for a critical discussion of the findings by Bitchener, 2008, and see Bitchener, 2009 for a response).

Research into the effectiveness of unfocused CF

As compared to the growing amount of evidence on the efficacy of focused CF, proof on the language learning potential of comprehensive or unfocused CF (i.e. correction of every error in students’ writing) is scarce. Notwithstanding the significant contribution of the focused CF work to the error correction debate, we still consider it important to explore the effect of unfocused CF on new pieces of writing. To begin with, the unfocused approach most closely resembles the correction method used in actual teaching practice; a teacher’s purpose in correcting his/her pupils’ written work is (among other things) to improve accuracy in general, not just the use of one grammatical feature (Ferris, 2010; Storch, 2010). This is why Hartshorn et al. (2010) called for research on a more authentic CF methodology, which focuses “on the accurate production of all aspects of writing, simultaneously” (p. 89). Moreover, Bruton (2009a) questioned the extent to which focused CF studies could still be considered to concern genuine L2 writing. He reasoned that, in just
focusing on one linguistic feature, their materials and CF rather seem to constitute written
grammar exercises instead of authentic writing tasks. Xu (2009) addressed a similar point
by stating that such a clear focus on one grammatical structure, may lead students to
consciously monitor the use of that target feature when performing the post-test(s). Finally,
the implications that can be drawn from focused CF studies so far are rather limited,
because the targeted linguistic feature (i.e. articles) was selected for maximal simplicity
(Ferris, 2010; Truscott, 2010).

Only two recent studies that we are aware of, aimed at investigating if unfocused or
comprehensive CF yields a learning effect (cf. Chapter 3; Truscott & Hsu, 2008). Truscott
and Hsu (2008) examined the writing performance of 47 ESL learners, half of whom
received comprehensive corrections while the other half functioned as a control group.
Truscott and Hsu found that, while comprehensive CF enabled their learners to improve the
accuracy of a particular text during revision, it did not lead to accuracy gains in a new text.
However, the fact that unfocused CF did not lead to learning in this study might have been
attributable to a ceiling effect (Bruton, 2009a); the texts learners wrote during the pre-test
held very few errors to begin with, and, as a result, little room was left for CF invoked
improvement in the post-test. The second study into the long-term effectiveness of
comprehensive CF was conducted as a pilot for the present study (cf. Chapter 3). It
explored the effects of two types of comprehensive CF and two control treatments on the
writing of 66 L2 learners of Dutch, and indicated that comprehensive error correction can be
beneficial to the SLA process. Findings showed that comprehensive error correction not
only led to improved accuracy in the revised version of a particular piece of writing, but that
it also yielded a learning effect; learners who received unfocused CF made significantly
fewer errors in newly produced texts than pupils whose errors had not been corrected.

Research into the relative effectiveness of focused and unfocused CF
Although hypotheses about the relative efficacy of focused and unfocused CF have been
forwarded in the literature, there is little empirical evidence on the differential value of the
two feedback methodologies. The only two studies addressing this issue are one by Ellis et
al. (2008), and one by Sheen, Wright, and Moldawa (2009). Ellis et al. (2008) did not find
any difference in accuracy gains between their focused and unfocused CF groups.
However, this study has several methodological weaknesses (see Xu, 2008 for a
discussion). One of the problems the authors themselves mentioned, is that students in the
focused group received more feedback on the target feature (i.e. articles) than students in
the unfocused group. Sheen et al. (2009) found the focused approach to be more beneficial
than provision of comprehensive feedback. However, as acknowledged by the authors themselves, the CF received by the unfocused group was rather unsystematic in nature; while some errors were corrected, others were ignored. It is conceivable that this unsystematic way of correcting has negatively influenced the effect of unfocused CF in this study. Students might have been confused noticing that some of their errors were disregarded.

4.3.2 Research into the relative effectiveness of direct and indirect CF
CF researchers have not only shown interest in the question if correction should be comprehensive or selective in nature. Many studies have also been exploring the relative effectiveness of different CF types. Most of the studies that compared different forms of CF categorized their CF methodologies as either direct or indirect. Whereas direct CF consists of an indication of the error and the corresponding correct L2 form, indirect CF only indicates that an error has been made. Instead of the teacher providing the target form, it is left up to the learner to correct his/her own errors. Indirect correction methods can take different forms that vary in their explicitness (e.g. underlining of errors, coding of errors).

Various hypotheses concerning the relative effectiveness of direct and indirect CF have been put forward, some in favor of direct error correction, others supporting the indirect approach. On the one hand it has been suggested that learners will benefit more from indirect CF because it engages students in a more profound form of language processing as they are self-editing their writing (e.g. Ferris, 1995; Lalande, 1982). In this view, the indirect approach “requires pupils to engage in guided learning and problem solving and, as a result, promotes the type of reflection that is more likely to foster long-term acquisition” (Bitchener & Knoch, 2008, p. 415). Advocates of direct CF (e.g. Chandler, 2003), on the other hand, claimed that the indirect approach might fail because indirect CF provides learners with insufficient information to resolve complex linguistic errors (e.g. syntactic errors). Chandler (2003) furthermore argued that, whereas direct CF enables learners to instantly internalize the correct form, learners whose errors are corrected

---

2 Sheen et al. (2009) justified the unsystematic nature of the unfocused CF treatment in their study by arguing that, in reality, teachers also tend to provide comprehensive feedback in an unsystematic and confusing manner. Even if this is the case, however, it would not validate doing the same for research purposes. What Sheen et al. (2009) have been comparing is not focused and unfocused CF per se, but focused CF and an inferior form of unfocused CF. The fact that the focused approach proved to be more beneficial than comprehensive correction in this study, does not warrant the conclusion that unfocused CF is necessarily less effective. It only shows that high quality focused feedback is more useful than badly provided unfocused CF.
indirectly do not know if their own hypothesized corrections are indeed accurate. This delay in access to the target form might level out the potential advantage of the additional cognitive effort associated with indirect CF. Moreover, it has been suggested that learners need a certain level of (meta-)linguistic competence to be able to self-correct their errors using indirect CF (e.g. Ferris, 2004; Hyland and Hyland, 2006; Sheen, 2007).

Neither of the above hypotheses could yet be confirmed by empirical findings, since clear evidence on the differential effects of direct and indirect CF on accuracy development is lacking. Research on the issue has produced conflicting results (cf. Chapter 2, section 2.5.2). Studies by Frantzen (1995) and Robb, Ross, and Shortreed (1986) failed to find any clear efficacy differences between direct and indirect types of correction. Ferris (2006) and Lalande (1982), on the other hand, reported an advantage of indirect over direct CF. However, Ferris’ study was not initially designed to directly compare the two CF methodologies, and in Lalande’s study, the observed between-group difference in accuracy gains was not statistically significant. Moreover, Lalande’s direct and indirect treatments differed in more respects than just the method of CF provision; the indirect group was engaged in more form-focused activities than the group receiving direct CF. Three other studies concluded that direct correction was the most effective approach (cf. Chapter 3; Bitchener & Knoch, 2010b; Chandler, 2003;). In Chandler’s (2003) study, however, the observed difference between direct CF and indirect correction was not significant. Additionally, the fact that students in this study received both direct and indirect CF in an only partial balanced design, makes it difficult – if not impossible – to come to any conclusions on the relative value of different CF types. In the study piloting the present research methodology (cf. Chapter 3), the difference between direct and indirect CF treatments did not reach significance either, at a p-value of .06. However, learners receiving direct CF significantly outperformed pupils in the control groups when writing a new text, whereas this study’s indirect CF group did not. Finally, Bitchener and Knoch’s (2010b) study did report a statistically significant difference between direct and indirect CF, in favor of the direct approach.

4.3.3 Research into the value of CF for different error types

In his 2007 article, Truscott explained that his case against CF (Truscott, 1996, 1999) was actually a case against grammar correction. He claimed that syntactic errors in particular might not be amenable to correction, because they are integral parts of a complex system which – in Truscott’s view – is impermeable to CF. He furthermore suggested that morphological features are evenly unlikely to benefit from CF because their acquisition not
only depends on the understanding of form, but also of meaning and use in relation to other words and portions of the language system. Truscott (2001; 2007) concluded that if CF has any value for L2 development, this could only be true for “errors that involve simple problems in relatively discrete items” (Truscott, 2001, p. 94) – such as spelling errors – and not for errors in grammar.

A number of studies (e.g. Bitchener et al., 2005; Ferris, 2006; Ferris & Roberts, 2001; Frantzen, 1995; Lalande, 1982; Sheppard, 1992) explored the effects of CF on separate error types, and all reported differing levels of improvement for different types of errors (cf. Chapter 2, section 2.5.3). None of these studies, however, could test Truscott’s claim that grammar correction is ineffective, because they did not explicitly and systematically investigate whether CF is more beneficial for non-grammatical error types than for errors in grammar.

4.3.4 Research into the potential harmful side-effects of CF

One of Truscott’s (2004; 2007) reasons to object against the use of CF in L2 writing classes, is that it may lead to simplified writing by triggering learners to avoid situations in which they make errors. These considerations led Truscott to propose that accuracy gains found in earlier correction studies might well be attributable to such avoidance and simplified writing instead of to CF. Truscott’s suggestions are in line with limited capacity models of attention which also predict a trade-off between accuracy and complexity (e.g. Skehan, 1998). Within these models L2 performance is expected to become more complex when learners are willing and feeling free to experiment with the target language. A focus on accuracy, on the other hand, “is seen to reflect a greater degree of conservatism” in which learners will try “to achieve greater control over more stable [interlanguage] elements” while avoiding extending their L2 repertoire (Skehan & Foster, 2001, p. 191).

Few studies have investigated the influence of written CF on linguistic complexity, and in our opinion, studies that did (Chandler, 2003; Robb et al., 1986; Sheppard, 1992), could not come to any warranted conclusions. Sheppard (1992), for example, reported a negative effect of CF on the structural complexity of learners’ writing, but in fact, his finding was non-significant. Robb et al. (1986) found CF to have a significant positive effect on written complexity, but this study did not include a control group who did not receive CF. The same holds for Chandler’s (2003) study, who did not find any effect of CF on the complexity of students’ writing. An additional problem with the latter study is that Chandler based her conclusion on holistic ratings. In our view, however, the fact that holistic ratings did not change does not necessarily prove that complexity did not change either. (See
A second argument that led Truscott (1996; 2004) to conclude that CF should be expected to harm L2 learners’ accuracy development, is that it diverts time and energy away from more productive aspects of writing instruction. The only study that directly tested this claim by comparing the effects of CF to those of writing practice, is one by Sheen et al. (2009). Their results opposed Truscott’s claim by showing that learners did not benefit more from writing practice than from CF.

4.4 Rationale of the present study and research questions

Based on the empirical evidence presented in the previous sections, the error correction debate cannot be fully settled. The present study – of which the set-up, tasks, and procedure were piloted (cf. Chapter 3) – intended to add to the existing body of research by trying to tackle some of the unsettled issues, and answering eight research questions, which will shortly be introduced.

A number of early CF studies showed that comprehensive or unfocused CF enables learners to improve a particular piece of writing through revision. Moreover, there is now growing evidence on the contribution of focused CF to the long-term acquisition of certain linguistic features. More empirical evidence is needed, however, before any definitive conclusions can be drawn about the value of unfocused or comprehensive CF for accuracy development or learning. Hence, the present study aimed to explore the value of comprehensive correction during revision, as well as its ability to yield a learning effect, while adopting the tightly controlled methodology of recent focused CF studies (e.g. Bitchener & Knoch, 2010a):

RQ 1 Is comprehensive written CF useful as an editing tool, in that it enables learners to improve the accuracy of an initial text during revision?

RQ 2 Does comprehensive written CF yield a learning effect, in that it leads to improved accuracy in new texts written one week and four weeks after CF has been provided?

Because it seems plausible that learners benefit from taking a critical look at their own text and revising it, even without teacher intervention, the present study furthermore opted to test if comprehensive CF has an added value above self-correction:
RQ 3 Is comprehensive CF more beneficial to learners’ accuracy development than having the opportunity to correct their own writing?

A second issue on which CF research to date was unable to come to any clear conclusions, is the relative efficacy of direct and indirect CF. The present study therefore set out to investigate the differential value of these two CF methodologies:

RQ 4 Which feedback methodology is most effective: direct or indirect CF?

Moreover, since no earlier research directly addressed Truscott’s (2001; 2007) claim that CF might have value for non-grammatical errors, but not for errors in grammar, we strived to test this hypothesis by distinguishing between grammatical and non-grammatical error types in our analyses:

RQ 5 Are grammatical errors less correctable than other types of errors (i.e. non-grammatical errors)?

In addition, very few studies have explored the potential harmful effects of CF. The present study therefore compared the effect of comprehensive correction to the effect of mere writing practice to determine if CF is a waste of time (Truscott, 1996; 2004). Moreover, we examined the lexical and structural complexity of learners’ output to see if pupils are inclined to avoid more complex structures due to error correction (Truscott, 2004; 2007):

RQ 6 Is comprehensive CF more beneficial to learners’ accuracy development than writing practice?
RQ 7 Does error correction lead to avoidance of lexically and structurally (more) complex utterances?

Finally, we explored the potential influence of learners’ educational level on the degree to which they are able to benefit from direct and indirect CF. As was explained in Chapter 2 (section 2.6.3), exploration of this issue is primarily interesting from a practical perspective, but may also lead to theoretical implications. In the first place, it would be valuable for teachers to know if learners from different educational levels are equally receptive to (direct and indirect) CF. Moreover, we presumed learners’ educational level to be indicative of their level of meta-linguistic awareness (cf. Chapter 2 for the rationale
behind this assumption). The hypothesis that learners with lower levels of (meta-)linguistic competence might be less able to correct their own errors based on indirect CF (e.g. Ferris, 2004; Hyland & Hyland, 2006; Sheen, 2007), then led us to expect that pupils with a higher educational level would profit more from indirect CF than pupils from a lower level of education:

RQ 8 What (if any) is the influence of pupils’ educational level on CF efficacy?

4.5 Methodology

4.5.1 Setting and participants

Four Dutch secondary schools with multilingual student populations participated in the study. Over 80% of those schools’ pupils came from non-Dutch language backgrounds; although most pupils were born in the Netherlands, many of them only started learning Dutch in school (i.e. at age four). Our sample was very heterogeneous with respect to language background (28 L1’s), Moroccan Arabic (31%), Turkish (16%), and Surinamese languages (16%: 8% Sranan Tongo, 8% Sarnami Hindustani) being the most common L1’s.

All schools that took part in our study aimed at integrating content and (second) language instruction by adopting a language sensitive instructional approach (e.g. Van Eerde & Hajer, 2005). Following this approach, language did not only play a central role in language classes, but also in classes whose overriding focus was on content (e.g. biology, mathematics, geography). The main aim of this language sensitive approach to content teaching is to cater for the special needs of L2 learners and learners with limited language proficiency, who might experience problems understanding and acquiring the content due to the linguistic demands of the input. The present investigation was conducted during biology classes, and our tasks treated biology-related topics.

The study’s population consisted of seven intact classes of pupils in their second year of secondary pre-vocational education (or vmbo-t in Dutch) (N = 134), and six classes of pupils in their second year of higher general secondary education (or havo in Dutch) (N = 134). In the remainder of this chapter we will use the contrast ‘lower-level’ and ‘higher-level’ to refer to these two groups of pupils respectively. Pupils mean age was 14 (min. 14, max. 15). Within classes participants were randomly assigned to the different treatment groups incorporated in the study, to prevent a confounding interaction between treatment and
class. Furthermore, the male/female and Dutch L1/L2 proportion\(^3\) was kept constant across treatment groups.

4.5.2 Treatments

The present study integrated two experimental treatments and two control treatments:

**Experimental group I: direct CF (DIR)**

Pupils in the first experimental group received comprehensive (or unfocused) direct CF. The researcher identified all existing linguistic errors, and provided the pupils with the corresponding target forms, as illustrated in example 1 below.

```
Example 1: direct error correction
De lieveheersbeestje is rood.          [The ladybug is red.]
Het lieveheersbeestje
```

**Experimental group II: indirect CF (IND)**

Texts of pupils in the second experimental group were corrected indirectly; the researcher provided an indication of each error and its category (cf. example 2), but it was left to the student to derive the corresponding target forms. We used nine different codes, classified under three coordinating categories, to correct different linguistic error types: (1) **lexical errors**: word choice; (2) **grammatical errors**: word form (e.g. verb tense, singular/plural), word order, incomplete sentences, and addition or omission of a word; and (3) **orthographical errors**: spelling, punctuation, and capitalization (cf. Appendix C).

---

\(^3\) We decided not to eliminate L1 speakers of Dutch from our sample because we consider ‘native-speakeress’ to be a sliding scale. Since the large majority of our participants was born in the Netherlands, it was unfeasible to decide which pupils might have had access to (some) Dutch language input from birth onwards and which learners did not. What is shared by all our participants, however, is a relatively low level of writing proficiency (i.e. as judged by their teachers). This is why CF was expected to be beneficial to all learners in our sample, irrespective of their language background. Nevertheless, we did re-run all analyses excluding the pupils for whom Dutch was the only language spoken at home (N = 24). Results from these analyses did not differ in any respect from those reported in this chapter.
CHAPTER 4

Control group I: self-correction (SLF)
Pupils in the first control group did not receive any feedback on their initial text but were invited to revise their writing without any available CF – that is to self-correct their output. Including this control treatment enabled us to set apart effects of error correction from effects of the revision process as such.

Control group II: additional writing practice (PRC)
Just as the first control group, participants in the second control group did not receive CF. Whereas pupils in the first control group were invited to revise their own text, however, pupils in the second control group were not involved in any revision activities. Instead they performed a completely new writing task, which provided them with the opportunity to practice their writing skills once more. Inclusion of this second control treatment assured an equal distribution of time-on-task across experimental and control treatments, since completion of a completely new task asked (at least) as much time as revising an already written text. Moreover, integration of the practice treatment enabled us to determine if CF is more effective than just providing pupils with more practice opportunities (Truscott, 1996; 2004).

4.5.3 Procedure
The present investigation included four sessions; a pre-test session (S1), a treatment session (S2), a post-test session (S3), and a delayed post-test session (S4). Figure 4.1 presents an overview of the study’s set-up. Our tasks were designed for research purposes only; they were not part of the standard biology curriculum. However, all tasks were administered during class periods. The tasks and topics were introduced and explained by the researcher, and the class teacher was present to maintain order. All tasks and tests were pen-and-paper assignments.

Example 2: indirect error correction

De lievenheersbeestje **S** is rood.  

([The ladybug is red.]

( __ = wrong word, **S** = spelling error)
Session 1: pre-tests (week 1)
During the first session (S1), all pupils (irrespective of their treatment group) were presented with a receptive vocabulary test (cf. Appendix E), a questionnaire concerning their language background (cf. Appendix D), and the first writing task. The output on this first assignment served as a baseline measure of pupils’ written accuracy and complexity. Pupils were given 20 minutes to complete the writing task, and were instructed to write a minimum of 15 lines. Scores on the vocabulary test were taken to provide an indication of pupils’ overall language proficiency. We included these scores in our analyses as a covariate to be able to control for individual L2 proficiency differences. A vocabulary test was chosen because earlier research demonstrated that vocabulary knowledge is a good predictor of overall language proficiency (e.g. Beglar & Hunt, 1999; Zareva, Schwanenflugel, & Nikolova, 2005). Moreover, a vocabulary test seemed to be the most suitable instrument out of practical considerations and time restrictions.

Before administering the first assignment, the researcher introduced the task’s topic by means of a ten minute mini-lesson, to ensure a comparable minimal amount of background knowledge on the topic amongst participants. Additionally, pupils were asked to focus on content as well as on form while writing. To give them an idea of the linguistic features they could attend to, pupils were given a hand-out listing common types of errors, and an example for each error category. We chose to draw learners’ attention to form at this stage, to be able to establish the unique contribution of CF in the end. If only the attention of the experimental groups would have been directed to linguistic form, it would have been impossible to unambiguously explain any apparent advantage for these groups over the control treatments; it would have been unclear if it was the form focus as such or the CF and/or revision that led to the effect. Moreover, the hand-out used here was re-used – in a slightly adapted version including error codes – when instructing pupils in the indirect CF group on interpreting the error codes (cf. section about session 2). This set-up ensured that the indirect group did not have an advantage as compared to the other groups in terms of language input. It is important to note that, whereas we directed pupils’ attention to language form, it was made clear that their main focus needed to be on the content of their writing. Furthermore, pupils were not aware of the fact that they could be receiving CF at a later stage.

Session 2: treatment (week 2)
The treatment session (S2) took place one week after S1. Pupils in the two CF groups received the corrected versions of their initial texts. They were asked to copy the initial
texts, revising all errors corrected by the researcher. Before starting their revision, pupils in the indirect CF group were given a hand-out (cf. Appendix C), accompanied by an oral instruction, on how to interpret and use the error codes in their texts. Pupils in the self-correction group were handed the texts they wrote during the pre-test session without any alterations. They were invited to self-correct their pre-test writing, by thoroughly reading over their texts and searching for any elements that needed revising. Even if no such utterances would be found, pupils were still obliged to copy their initial texts. (Cf. Appendix B for the instructions received by the direct CF, indirect CF, and self-correction groups.) Finally, pupils in the practice group were not given the opportunity to revise their pre-test texts, but were presented with a new writing task instead. The researcher shortly explained the new task’s topic before pupils started writing. All pupils were allocated 20 minutes to finish the task they were presented with, irrespective of the treatment group they were assigned to.

Sessions 3 and 4: post-test (week 3) and delayed post-test (week 6)
The first post-test (S3) was administered one week after the treatment session (S2), and the delayed post-test took place four weeks after S2. During both post-test sessions, all pupils were given 20 minutes to produce a text (at least 15 lines in length) on a new topic, which was again shortly introduced by the researcher.

Figure 4.1 Experimental set-up

<table>
<thead>
<tr>
<th>Treatment group</th>
<th>Pre-tests</th>
<th>Treatment</th>
<th>Post-test</th>
<th>Delayed post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1: week 1</td>
<td>S2: week 2</td>
<td>S3: week 3</td>
<td>S4: week 6</td>
</tr>
<tr>
<td>Direct CF (DIR)</td>
<td>Vocabulary test &amp; Initial text</td>
<td>revision based on comprehensive direct CF</td>
<td>New text</td>
<td>New text</td>
</tr>
<tr>
<td>Indirect CF (IND)</td>
<td></td>
<td></td>
<td>revision based on comprehensive indirect CF</td>
<td></td>
</tr>
<tr>
<td>Self-Correction (SLF)</td>
<td></td>
<td></td>
<td>revision based on self-correction (no CF)</td>
<td></td>
</tr>
<tr>
<td>Practice (PRC)</td>
<td></td>
<td></td>
<td>writing practice (no revision or CF)</td>
<td></td>
</tr>
</tbody>
</table>
4.5.4 Writing tasks

The tasks used throughout the study were four writing assignments on a biology-related topic, namely the metamorphosis of different insects (i.e. butterflies (S1), honeybees (S2), ladybugs (S3), and wasps (S4)). Whereas three of the tasks were performed by all pupils, the task on a wasp's metamorphosis was only performed by participants in the second control group, who practiced their writing during the treatment session (S2).

All tasks within the series had a comparable form and set-up; they invited pupils to write an e-mail to a classmate who was absent during the researcher’s explanation of the task’s topic. Pupils were asked to explain the metamorphosis of the relevant insect, based on a series of images depicting the metamorphosis process. To avoid individual differences in familiarity with the relevant vocabulary, (potentially) unfamiliar words (e.g. larva, cocoon) were glossed. (Cf. Appendix A.)

4.5.5 Data processing

All hand written output was transcribed and coded using CLAN (Computerized Language ANalysis) software, a program designed to analyze data transcribed in the CHILDES (Child Language Data Exchange System) format (MacWhinney, 2000). Two research assistants transcribed pupils’ texts, while the first author was responsible for coding them. Texts were coded for linguistic errors and clause types (i.e. main clauses and subordinate clauses). The coding procedure was entirely blind; during coding the researcher was unaware of the treatment group the text at hand belonged to. Ten percent of the data was also coded by one of the assistants to be able to establish inter-rater reliability, and re-coded, six months later, by the first author to measure intra-rater reliability.

We calculated intra-class correlation coefficients (ICC) to establish the average levels (over sessions) of intra-rater and inter-rater agreement for overall accuracy, grammatical accuracy, non-grammatical accuracy, and structural complexity (cf. section 4.5.6). The intra-rater ICC was calculated from an ANOVA two-way mixed effects model, and provides an indication of the variability due to variation within the same rater. A two-way random effects model was used to estimate the inter-rater ICC, reporting the proportion of the variability due to variation among raters. Results show high levels of agreement within the same rater as well as between raters for all four measures (cf. Table 4.1).

---

4 Since we only compared accuracy performance across groups and not sessions, tasks were not counterbalanced.
Table 4.1 Average levels of intra-rater and inter-rater agreement for measures of accuracy and structural complexity

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>Grammatical accuracy</th>
<th>Non-grammatical accuracy</th>
<th>Structural Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC intra-rater</td>
<td>.998</td>
<td>.977</td>
<td>.996</td>
<td>.974</td>
</tr>
<tr>
<td>ICC inter-rater</td>
<td>.981</td>
<td>.921</td>
<td>.975</td>
<td>.946</td>
</tr>
</tbody>
</table>

4.5.6 Linguistic measures

Every text was analyzed for accuracy, structural complexity, and lexical diversity. As did earlier studies exploring the effectiveness of written CF (e.g. Chandler, 2003; Truscott & Hsu, 2008), we used an error ratio to measure overall accuracy (i.e. (number of form-related errors/total number of words) x 10). A ten-word ratio rather than the more common 100-word ratio was used because pupils’ texts were relatively short (i.e. around 120 words).

To be able to test Truscott’s (2001; 2007) claim that non-grammatical errors might be more correctable than errors in grammar, we broke down our overall accuracy measure into a measure of grammatical accuracy and a measure of non-grammatical accuracy. The first measure was a ratio calculated on the basis of the sum of the number of article errors, inflectional errors, word order errors, omissions of a necessary element, additions of a non-necessary element, pronominal errors, and other grammatical errors (i.e. (number of grammatical errors/total number of words) x 10). Lexical errors, orthographical errors, appropriateness/pragmatic errors, and other non-grammatical errors, on the other hand, were included in the non-grammatical accuracy ratio (i.e. (number of non-grammatical errors/total number of words) x 10).

To allow investigation of the possible influence of error correction on the complexity of pupils’ writing, we analyzed all texts for both structural complexity and lexical diversity. We used a subordination index, that is the number of subordinate clauses as a percentage of the total number of clauses (i.e. (number of sub clauses/total number of clauses) x 100), to obtain an indication of a text’s structural complexity (Norris & Ortega, 2009; Wolfe-

---

5 Although we were able to classify almost all of the observed errors, some did not fit in either of the specified error subcategories. These were brought together under ‘other grammatical errors’ or ‘other non-grammatical errors’.

6 The error categories we distinguished by different error codes when providing indirect CF (cf. Appendix C), are not in a one-to-one relation with the categories used during the coding procedure. For example, we used the same code – that is ‘wrong word’ – for article errors, pronominal errors, and lexical errors when providing indirect CF to keep the code labels as transparent as possible for our learners.
Quintero, Inagaki, & Kim, 1998). Lexical diversity was calculated using Guiraud’s Index, a type-token ratio that corrects for text length (types/√tokens) (Guiraud, 1954).

4.5.7 Analysis
As our participants came from different classes within different schools, our data were structured hierarchically. The performances of pupils within one class or school were expected to be more similar to each other than to performances of pupils in other classes and schools; a teacher, for example, is likely to affect the performances of his/her pupils. If observations within classes or schools had indeed been dependent, this would have affected the choice of appropriate statistical procedures. Analysis of variance for instance, relies on the independency of cases. We applied linear multilevel analyses (e.g. Snijders & Bosker, 1999) to explore the relationship between observations within classes and schools. The multilevel procedure enabled us to explicitly model possible dependencies in the data by including class and school as random factors. These multilevel analyses showed, however, that class and school never made a significant contribution to the model. After we had ascertained that the assumption of independence was met, we proceeded analyzing our data using AN(C)OVA’s because this procedure has proven to have more power in absence of data dependencies (e.g. Snijders & Bosker, 1999).

We used ANCOVA’s to test for between-group differences on the different dependent variables (i.e. overall accuracy, grammatical accuracy, non-grammatical accuracy, lexical diversity, and structural complexity) in the treatment session (S2), first post-test session (S3), and delayed post-test session (S4). The initial ANCOVA models contained treatment and educational level as between-subject variables, and language proficiency as a covariate. In addition, we incorporated learners’ pre-test (S1) performance on the relevant dependent variable as a covariate to account for effects of initial individual differences. We started out each ANCOVA with a full model, including all relevant factors, and all possible interactions between those factors. (Interactions between) factors, which did not explain a significant proportion of the variance, were excluded from the final ANCOVA models.

---

7 Because we only have two to four observations (i.e. classes) per school in our sample, we never included both random factors class and school in the same analysis. Instead we performed two separate analyses to find out if any of those contextual variables brought about unplanned dependencies in the data.
4.6 Results

In this section we will first present some relevant descriptive statistics, and describe learners' performance on the writing and language proficiency pre-tests. We will then successively present the findings regarding the value of comprehensive CF as an editing tool (RQ 1), its language learning potential (RQ 2), the effect of direct and indirect CF on grammatical and non-grammatical issues (RQ’s 4 and 5), and the influence of CF on the complexity of learners' writing (RQ 7). The comparison between the effects of CF and those of self-correction (RQ 3) and additional writing practice (RQ 7), and the role of pupils’ educational level (RQ 8), will be discussed throughout the results section.

4.6.1 Descriptive statistics

Language proficiency

The descriptive statistics for overall language proficiency (i.e. the score on a receptive vocabulary test), itemized per educational level and treatment group are presented in Table 4.2.

Table 4.2 Descriptive statistics: overall language proficiency\(^a\) by educational level and treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Language proficiency (M)</th>
<th>SD</th>
<th>Lower educational level (vmbo-t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIR (N = 32)</td>
<td>76.09</td>
<td>10.52</td>
<td>DIR (N = 34)</td>
</tr>
<tr>
<td>IND (N = 26)</td>
<td>74.85</td>
<td>11.80</td>
<td>IND (N = 31)</td>
</tr>
<tr>
<td>SLF (N = 33)</td>
<td>80.45</td>
<td>8.93</td>
<td>SLF (N = 31)</td>
</tr>
<tr>
<td>PRC (N = 33)</td>
<td>74.09</td>
<td>8.46</td>
<td>PRC (N = 33)</td>
</tr>
<tr>
<td>Total (N = 124)</td>
<td>76.46</td>
<td>10.10</td>
<td>Total (N = 129)</td>
</tr>
</tbody>
</table>

\(^a\)Number of correct items (out of 108) on receptive vocabulary test.

Accuracy

Table 4.3 displays the descriptive statistics for overall accuracy scores for all treatment groups, itemized per educational level and session (i.e. pre-test, treatment session, post-
test, and delayed post-test). Tables 4.4 and 4.5 provide the descriptives for grammatical and non-grammatical accuracy respectively.

**Table 4.3** Descriptive statistics: overall accuracy\(^a\) by educational level, treatment, and session

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Treatment</th>
<th>Pre-test (S1)</th>
<th>Treatment session (S2)</th>
<th>Post-test (S3)</th>
<th>Delayed post-test (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>High (havo)</td>
<td>DIR (N = 32)</td>
<td>1.52 0.75</td>
<td>0.37 0.29</td>
<td>1.08 0.66</td>
<td>1.32 0.89</td>
</tr>
<tr>
<td></td>
<td>IND (N = 29)</td>
<td>1.25 0.67</td>
<td>0.42 0.32</td>
<td>1.05 0.46</td>
<td>0.89 0.40</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 37)</td>
<td>1.39 0.60</td>
<td>1.28 0.66</td>
<td>1.61 0.82</td>
<td>1.50 0.70</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 36)</td>
<td>1.29 0.63</td>
<td>1.30 0.72</td>
<td>1.49 0.73</td>
<td>1.41 0.77</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>1.37 0.66</td>
<td>0.88 0.71</td>
<td>1.33 0.73</td>
<td>1.31 0.75</td>
</tr>
<tr>
<td>Low (vmbo-t)</td>
<td>DIR (N = 35)</td>
<td>1.69 0.70</td>
<td>0.32 0.24</td>
<td>1.64 0.76</td>
<td>1.42 0.78</td>
</tr>
<tr>
<td></td>
<td>IND (N = 33)</td>
<td>1.95 0.88</td>
<td>0.74 0.63</td>
<td>1.73 0.97</td>
<td>1.54 0.90</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 34)</td>
<td>1.85 0.75</td>
<td>1.50 0.68</td>
<td>1.97 0.93</td>
<td>1.75 0.84</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 32)</td>
<td>1.79 1.00</td>
<td>2.19 1.00</td>
<td>2.26 1.00</td>
<td>2.08 1.10</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>1.82 0.83</td>
<td>1.17 0.99</td>
<td>1.90 0.94</td>
<td>1.70 0.93</td>
</tr>
</tbody>
</table>

\(^a\)Number of form-related errors per 10 words.

**Table 4.4** Descriptive statistics: grammatical accuracy\(^a\) by educational level, treatment, and session

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Treatment</th>
<th>Pre-test (S1)</th>
<th>Treatment session (S2)</th>
<th>Post-test (S3)</th>
<th>Delayed post-test (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>High (havo)</td>
<td>DIR (N = 32)</td>
<td>0.43 0.27</td>
<td>0.09 0.10</td>
<td>0.32 0.22</td>
<td>0.39 0.34</td>
</tr>
<tr>
<td></td>
<td>IND (N = 29)</td>
<td>0.30 0.24</td>
<td>0.10 0.11</td>
<td>0.36 0.27</td>
<td>0.37 0.22</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 37)</td>
<td>0.37 0.25</td>
<td>0.37 0.23</td>
<td>0.42 0.30</td>
<td>0.46 0.33</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 36)</td>
<td>0.37 0.27</td>
<td>0.36 0.27</td>
<td>0.44 0.23</td>
<td>0.45 0.23</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>0.37 0.26</td>
<td>0.24 0.24</td>
<td>0.39 0.26</td>
<td>0.42 0.29</td>
</tr>
<tr>
<td>Low (vmbo-t)</td>
<td>DIR (N = 35)</td>
<td>0.55 0.31</td>
<td>0.10 0.11</td>
<td>0.49 0.33</td>
<td>0.39 0.26</td>
</tr>
<tr>
<td></td>
<td>IND (N = 33)</td>
<td>0.47 0.33</td>
<td>0.21 0.23</td>
<td>0.51 0.27</td>
<td>0.49 0.32</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 34)</td>
<td>0.53 0.32</td>
<td>0.46 0.29</td>
<td>0.64 0.41</td>
<td>0.60 0.39</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 32)</td>
<td>0.52 0.36</td>
<td>0.59 0.41</td>
<td>0.69 0.44</td>
<td>0.72 0.46</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>0.52 0.33</td>
<td>0.34 0.34</td>
<td>0.58 0.38</td>
<td>0.55 0.38</td>
</tr>
</tbody>
</table>

\(^a\)Number of grammatical errors per 10 words.
### Table 4.5 Descriptive statistics: non-grammatical accuracy by educational level, treatment, and session

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Treatment</th>
<th>Pre-test (S1)</th>
<th>Treatment session (S2)</th>
<th>Post-test (S3)</th>
<th>Delayed post-test (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
</tr>
<tr>
<td>High (havo)</td>
<td>DIR (N = 32)</td>
<td>1.09 0.63</td>
<td>0.28 0.27</td>
<td>0.77 0.56</td>
<td>0.93 0.82</td>
</tr>
<tr>
<td></td>
<td>IND (N = 29)</td>
<td>0.95 0.52</td>
<td>0.32 0.26</td>
<td>0.69 0.39</td>
<td>0.52 0.29</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 37)</td>
<td>1.01 0.55</td>
<td>0.92 0.56</td>
<td>1.20 0.69</td>
<td>1.04 0.59</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 36)</td>
<td>0.93 0.54</td>
<td>0.94 0.63</td>
<td>1.05 0.66</td>
<td>0.97 0.69</td>
</tr>
<tr>
<td></td>
<td>Total (N = 134)</td>
<td>1.00 0.56</td>
<td>0.64 0.56</td>
<td>0.95 0.63</td>
<td>0.89 0.66</td>
</tr>
<tr>
<td>Low (vmbo-t)</td>
<td>DIR (N = 35)</td>
<td>1.14 0.56</td>
<td>0.23 0.17</td>
<td>1.15 0.65</td>
<td>1.03 0.75</td>
</tr>
<tr>
<td></td>
<td>IND (N = 33)</td>
<td>1.48 0.72</td>
<td>0.52 0.47</td>
<td>1.22 0.82</td>
<td>1.05 0.72</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 34)</td>
<td>1.32 0.63</td>
<td>1.05 0.64</td>
<td>1.33 0.69</td>
<td>1.15 0.70</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 32)</td>
<td>1.28 0.82</td>
<td>1.60 0.85</td>
<td>1.56 0.89</td>
<td>1.36 0.95</td>
</tr>
<tr>
<td></td>
<td>Total (N = 134)</td>
<td>1.30 0.69</td>
<td>0.83 0.78</td>
<td>1.32 0.78</td>
<td>1.15 0.79</td>
</tr>
</tbody>
</table>

*aNumber of non-grammatical errors per 10 words.*

Figure 4.2 is a graphic illustration of the descriptives presented in Table 4.3 (collapsed over educational level). It shows the accuracy development of the different treatment groups over time. The graph provides a descriptive preview of this study’s main findings, which will be presented in the following section. As can be seen from the graph, all groups performed similar at the pre-test (S1). In the treatment session (S2), however, the error rate of the CF groups and the self-correction group decreased, whereas the number of errors committed by the practice group increased. In the two post-tests (S3 and S4), we still see an error rate difference between the control and experimental groups, in favor of the latter.
Figure 4.2 Overall accuracy per treatment and session

![Overall accuracy chart](chart.png)
CHAPTER 4

Complexity
Table 4.6 shows the descriptives for structural complexity per treatment group, itemized per educational level and session. Lexical diversity scores are presented in Table 4.7.

Table 4.6 Descriptive statistics: structural complexity\(^a\) by educational level, treatment, and session

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Treatment</th>
<th>Pre-test (S1)</th>
<th>Treatment session (S2)</th>
<th>Post-test (S3)</th>
<th>Delayed post-test (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M SD</td>
<td>M SD</td>
<td>M SD</td>
<td>M SD</td>
</tr>
<tr>
<td>High (havo)</td>
<td>DIR (N = 32)</td>
<td>16.31 10.46</td>
<td>17.30 10.87</td>
<td>14.67 9.14</td>
<td>10.27 8.13</td>
</tr>
<tr>
<td></td>
<td>IND (N = 29)</td>
<td>16.30 8.98</td>
<td>18.07 10.20</td>
<td>14.36 10.38</td>
<td>10.98 8.25</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 37)</td>
<td>14.74 8.94</td>
<td>14.39 9.30</td>
<td>15.65 10.92</td>
<td>10.94 10.62</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 36)</td>
<td>19.86 10.02</td>
<td>16.13 8.73</td>
<td>15.75 9.32</td>
<td>15.80 8.02</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>16.83 9.71</td>
<td>16.35 9.74</td>
<td>15.16 9.88</td>
<td>12.06 9.10</td>
</tr>
<tr>
<td>Low (vmbo-t)</td>
<td>DIR (N = 35)</td>
<td>20.78 8.17</td>
<td>21.66 8.27</td>
<td>18.65 10.77</td>
<td>15.17 9.64</td>
</tr>
<tr>
<td></td>
<td>IND (N = 33)</td>
<td>19.41 11.66</td>
<td>21.04 10.57</td>
<td>16.70 10.98</td>
<td>10.84 8.95</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 34)</td>
<td>15.03 8.95</td>
<td>16.19 8.62</td>
<td>14.57 11.00</td>
<td>11.25 7.67</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 32)</td>
<td>18.86 8.00</td>
<td>15.19 8.23</td>
<td>13.66 8.65</td>
<td>13.17 8.30</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>18.52 9.45</td>
<td>18.57 9.32</td>
<td>15.89 10.45</td>
<td>12.72 8.73</td>
</tr>
</tbody>
</table>

\(^a\)Subordination Index (i.e. (number of sub clauses/total number of clauses) \times 100)

Table 4.7 Descriptive statistics: lexical diversity\(^a\) by educational level, treatment, and session

<table>
<thead>
<tr>
<th>Educational level</th>
<th>Treatment</th>
<th>Pre-test (S1)</th>
<th>Treatment session (S2)</th>
<th>Post-test (S3)</th>
<th>Delayed post-test (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M SD</td>
<td>M SD</td>
<td>M SD</td>
<td>M SD</td>
</tr>
<tr>
<td>High (havo)</td>
<td>DIR (N = 32)</td>
<td>7.13 0.85</td>
<td>7.28 0.91</td>
<td>6.87 0.90</td>
<td>6.51 0.80</td>
</tr>
<tr>
<td></td>
<td>IND (N = 29)</td>
<td>7.10 0.77</td>
<td>7.10 0.73</td>
<td>6.69 0.68</td>
<td>6.61 0.63</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 37)</td>
<td>7.09 0.81</td>
<td>7.06 0.78</td>
<td>6.70 0.90</td>
<td>6.31 0.98</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 36)</td>
<td>7.31 0.50</td>
<td>6.72 0.72</td>
<td>6.88 0.71</td>
<td>6.63 0.83</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>7.16 0.74</td>
<td>7.03 0.81</td>
<td>6.79 0.80</td>
<td>6.51 0.84</td>
</tr>
<tr>
<td>Low (vmbo-t)</td>
<td>DIR (N = 35)</td>
<td>7.00 0.81</td>
<td>7.26 0.70</td>
<td>6.51 0.73</td>
<td>6.58 0.69</td>
</tr>
<tr>
<td></td>
<td>IND (N = 33)</td>
<td>6.77 1.06</td>
<td>7.03 1.00</td>
<td>6.64 1.00</td>
<td>6.47 0.89</td>
</tr>
<tr>
<td></td>
<td>SLF (N = 34)</td>
<td>7.08 0.75</td>
<td>7.12 0.82</td>
<td>6.46 0.70</td>
<td>6.37 0.75</td>
</tr>
<tr>
<td></td>
<td>PRC (N = 32)</td>
<td>6.96 0.87</td>
<td>6.53 0.77</td>
<td>6.59 0.87</td>
<td>6.57 0.70</td>
</tr>
<tr>
<td>Total (N = 134)</td>
<td></td>
<td>6.95 0.88</td>
<td>7.00 0.87</td>
<td>6.55 0.74</td>
<td>6.50 0.75</td>
</tr>
</tbody>
</table>

\(^a\)Guiraud’s Index (types/\(\sqrt{\text{tokens}}\))
4.6.2 Pre-test performance

A series of ANOVA’s on pupils’ pre-test texts (S1) showed that there were no initial differences between treatment groups in overall accuracy ($F(3, 264) < 1, p = .902, \eta^2_p = .00$), grammatical accuracy ($F(3, 264) = 1.30, p = .276, \eta^2_p = .02$), non-grammatical accuracy ($F(3, 264) < 1, p = .620, \eta^2_p = .01$), and lexical diversity ($F(3, 264) < 1, p = .484, \eta^2_p = .01$). We did find a significant initial difference on our measure of structural complexity ($F(3, 264) = 3.04, p = .029, \eta^2_p = .03$). Post-hoc pair wise comparisons revealed that the pre-test writing of pupils in the practice group was structurally more complex than that of learners in the self-correction group ($p = .033$). Finally, we did not find a significant difference between treatment groups in language proficiency ($F(3, 253) = 1.96, p = .120, \eta^2_p = .02$).

Educational level turned out to be a significant predictor of language proficiency ($F(1, 251) = 19.15, p < .001, \eta^2_p = .07$); pupils with a higher educational level scored higher on the vocabulary test than pupils with a lower level of education.

Moreover, results did display a significant effect of educational level and language proficiency on learners’ pre-test writing. We found that both educational level ($F(1, 250) = 14.20, p < .001, \eta^2_p = .05$) and language proficiency ($F(1, 250) = 15.45, p < .001, \eta^2_p = .06$) predicted the overall accuracy of pupils’ S1 texts in such a way that the lower-level pupils committed significantly more errors than pupils with a higher level of education and language proficiency. The same holds for grammatical accuracy (educational level: $F(1, 250) = 10.83, p = .001, \eta^2_p = .04$; language proficiency: $F(1, 250) = 12.69, p < .001, \eta^2_p = .05$) and non-grammatical accuracy (educational level: $F(1, 250) = 8.81, p = .003, \eta^2_p = .03$; language proficiency: $F(1, 250) = 9.21, p = .003, \eta^2_p = .04$). Finally, the lexical diversity (but not the structural complexity) of pupils’ pre-test writing co-varied significantly with their language proficiency ($F(1, 250) = 7.98, p = .005, \eta^2_p = .03$).

4.6.3 Effects of comprehensive CF on written accuracy

As mentioned in section 4.5.7, ANCOVA’s were used to test for between-group accuracy differences in the treatment session (S2), post-test (S3), and delayed post-test (S4). The initial ANCOVA models contained treatment and educational level as between-subject variables, and language proficiency and overall accuracy S1 as covariates. Insignificant (interactions between) factors were excluded from the final models.
Revision effects (S2)

During revision, pupils’ language proficiency \( F(1, 240) = 2.63, p = .106, \eta^2_p = .01 \) and the interaction between treatment and language proficiency \( F(3, 240) = 2.40, p = .069, \eta^2_p = .03 \) did not prove to play a significant role. These factors were therefore discarded from the final ANCOVA model. The definitive analysis revealed significant effects of treatment \( F(3, 259) = 111.24, p < .001, \eta^2_p = .56 \) and educational level \( F(1, 259) = 4.29, p = .039, \eta^2_p = .02 \). We also found that the interaction between treatment and educational level made a significant contribution to the model \( F(3, 259) = 7.85, p < .001, \eta^2_p = .08 \). Furthermore, accuracy in the treatment session (S2) co-varied significantly with pre-test accuracy scores (S1) \( F(1, 259) = 112.77, p < .001, \eta^2_p = .30 \). When comparing the sizes of the different significant effects, it is interesting to note that the treatment learners received and their performance on the pre-test text (S1), have a large effect on the treatment session (S2) outcomes. Their educational level, on the other hand, only rendered a medium sized effect.

To be able to determine where the significant differences in overall accuracy lay between treatment groups, we conducted post-hoc pair wise comparisons, using Bonferroni adjustments. Table 4.8 presents an overview of all significant contrasts between treatments per session and accuracy type. The most important observation is that pupils were able to use the CF they received during revision; both CF treatments turned out to be significantly more beneficial than either of the control treatments. Pupils who received direct CF outperformed pupils in the self-correction \( p < .001, \text{Cohen's } d = 1.99 \) and practice group \( p < .001, \text{Cohen's } d = 2.74 \). Indirect CF also proved to be more constructive than both correcting without feedback \( p < .001, \text{Cohen's } d = 1.55 \) and practicing writing \( p < .001, \text{Cohen's } d = 2.29 \). A comparison between the two experimental groups did not reach significance. However, the Cohen’s \( d \) values show that the effects of direct CF were larger than those of indirect CF\(^9\). This was to be expected, because pupils receiving direct CF only needed to copy the target forms as provided by the researcher. We furthermore did observe significant differences between the two control treatments, which can be explained from their interaction with pupils’ educational level. Whereas the higher-level pupils benefited equally from practicing their writing and self-correcting their texts, self-correction had an advantage over writing practice for the lower-level pupils \( p < .001, \text{Cohen’s } d = 1.23 \) during the treatment session.

---

\(^8\) We applied the following rule of thumb in interpreting \( \eta^2_p \) values: small \( 0.01 \leq \eta^2_p < 0.06 \), medium \( 0.06 \leq \eta^2_p < 0.14 \), large \( \eta^2_p \geq 0.14 \).

\(^9\) According to Cohen (1988), effect sizes can be classified as either small (0.20), medium (0.50), or large (0.80).
**Table 4.8 Summary of significant contrasts between treatment groups**

<table>
<thead>
<tr>
<th></th>
<th>Overall accuracy</th>
<th>Grammatical accuracy</th>
<th>Non-grammatical accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment session (S2)</td>
<td>***DIR &gt; SLF</td>
<td>***DIR &gt; SLF</td>
<td>***DIR &gt; SLF</td>
</tr>
<tr>
<td></td>
<td>***DIR &gt; PRC</td>
<td>***DIR &gt; PRC</td>
<td>***DIR &gt; PRC</td>
</tr>
<tr>
<td></td>
<td>***IND &gt; SLF</td>
<td>***IND &gt; SLF</td>
<td>***IND &gt; SLF</td>
</tr>
<tr>
<td></td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
</tr>
<tr>
<td>Post-test (S3)</td>
<td>***DIR &gt; SLF</td>
<td>*DIR &gt; PRC</td>
<td>*DIR &gt; SLF</td>
</tr>
<tr>
<td></td>
<td>***DIR &gt; PRC</td>
<td>**DIR &gt; PRC</td>
<td>**DIR &gt; PRC</td>
</tr>
<tr>
<td></td>
<td>***IND &gt; SLF</td>
<td>**IND &gt; SLF</td>
<td>**IND &gt; SLF</td>
</tr>
<tr>
<td></td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
</tr>
<tr>
<td>Delayed post-test (S4)</td>
<td>**DIR &gt; PRC</td>
<td>**DIR &gt; PRC</td>
<td>**IND &gt; SLF</td>
</tr>
<tr>
<td></td>
<td>**IND &gt; SLF</td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
</tr>
<tr>
<td></td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
<td>***IND &gt; PRC</td>
</tr>
</tbody>
</table>

*p<.05, **p<.01, ***p<.001.

**Short-term learning effects (S3)**

The initial ANCOVA model applied to the data of our first post-test (i.e. the texts written one week after pupils received the different treatments) revealed that the interactions between treatment and educational level ($F(3, 225) < 1$, $p = .519$, $\eta_p^2 = .01$), and treatment and language proficiency ($F(3, 225) < 1$, $p = .436$, $\eta_p^2 = .01$) were not significant. These factors were therefore not included in the definitive ANCOVA model. Our final analysis showed that the four treatment groups significantly varied in the number of errors they made in the first post-test ($F(3, 231) = 11.34$, $p < .001$, $\eta_p^2 = .13$). Again, we found that pupils with a higher level of education outperformed those with a lower educational level ($F(1, 231) = 4.45$, $p = .036$, $\eta_p^2 = .02$). Moreover, S3 accuracy scores significantly co-varied with both language proficiency ($F(1, 231) = 12.93$, $p < .001$, $\eta_p^2 = .05$), and overall accuracy S1 ($F(1, 231) = 87.90$, $p < .001$, $\eta_p^2 = .28$). As was the case in the treatment session (S2), especially the effects of treatment and pre-test performance are noteworthy. Compared to the effects of these two factors, the effects associated with learners’ educational level and language proficiency are relatively small.

In the light of the treatment effect that was found, we performed post-hoc pair wise comparisons, using Bonferroni adjustments (cf. Table 4.8). We observed that both CF groups outperformed the two control groups. Pupils who received direct CF on the first piece of writing, made fewer errors in the new text they wrote a week later (S3), than pupils who self-corrected their errors during revision ($p = .002$, Cohen’s $d = 0.67$) or practiced their writing skills without any revising opportunity ($p < .001$, Cohen’s $d = 0.81$). Pupils whose
errors were corrected indirectly also performed significantly better on the first post-test than pupils in the self-correction ($p = .001$, Cohen’s $d = 0.70$) and practice ($p < .001$, Cohen’s $d = 0.84$) groups. There were no significant differences amongst the CF treatments; direct and indirect CF showed to be equally effective in improving pupils’ written accuracy (when the new text was written a week after the moment of feedback provision). When comparing the sizes of the S3 CF effects to the Cohen’s $d$ values reported in the previous section, we see that CF has a smaller effect on the accuracy of new texts than on revisions. To conclude, no significant accuracy differences were found between the two control groups.

**Long-term learning effects (S4)**

As in the first post-test session (S3), the interactions between treatment and educational level ($F(3, 205) < 1, p = .649, \eta^2_p = .01$), and treatment and language proficiency ($F(3, 205) < 1, p = .952, \eta^2_p = .00$), turned out to be non-significant during the delayed post-test (S4), which took place four weeks after the treatment session. Moreover, pupils’ educational level was no longer an explaining factor ($F(1, 205) < 1, p = .436, \eta^2_p = .00$). These non-significant parameters were deleted from the ANCOVA model. The final model revealed that there was a significant between-group difference concerning overall accuracy in the texts learners wrote four weeks after they received CF, self-corrected their initial text, or practiced their writing skills ($F(3, 212) = 8.45, p < .001, \eta^2_p = .11$). As on the first post-test, language proficiency ($F(1, 212) = 5.56, p = .019, \eta^2_p = .02$) and overall accuracy S1 ($F(1, 212) = 86.38, p < .001, \eta^2_p = .29$) proved to be significant covariates. Once more, the effect of learners’ language proficiency was much smaller than those of treatment and pupils’ performance on the pre-test (S1).

Bonferroni adjusted post-hoc pairwise comparisons revealed that pupils whose errors were corrected four weeks before writing a new text, still benefited from the CF they received (cf. Table 4.8). Both the direct ($p = .005$, Cohen’s $d = 0.63$) and indirect ($p < .001$, Cohen’s $d = 0.94$) groups made significantly fewer errors than pupils who practiced their writing skills during the treatment session (S2). Table 4.8 furthermore shows that pupils who received indirect CF also outperformed the group who corrected their own errors without any available feedback ($p = .005$, Cohen’s $d = 0.66$). Interestingly, the post-test (S3) and delayed post-test (S4) effects are comparable in size. This suggests that the CF invoked accuracy gains are not transient in nature.
4.6.4 Effects of direct and indirect CF on grammatical and non-grammatical errors

To test Truscott’s (2007) claim that CF could have value for non-grammatical errors but not for errors in grammar, we performed separate analyses for grammatical and non-grammatical error types. The treatment session (S2), post-test (S3), and delayed post-test (S4) outcomes of two series of ANCOVA’s are reported; one for grammatical accuracy and one for non-grammatical accuracy. The initial ANCOVA models contained treatment and educational level as between-subject variables, and language proficiency and (non-)grammatical accuracy S1 as covariates. To increase the clarity of our findings, this section only reports the observed significant (interactions between) main effects.

Revision effects (S2)

During revision, the treatment learners received proved to have a significant effect on both pupils’ grammatical accuracy \( (F(3, 259) = 51.09, p < .001, \eta^2_p = .37) \) and non-grammatical accuracy \( (F(3, 259) = 81.80, p < .001, \eta^2_p = .49) \). The outcomes of post-hoc pair wise comparisons with Bonferroni adjustments per accuracy type, showed that CF enabled learners to resolve both their grammatical and non-grammatical errors when revising their initial text (cf. Table 4.8). Pupils who received direct CF outperformed learners in the self-correction group on both grammatical \( (p < .001, \text{Cohen's } d = 1.52) \) and non-grammatical \( (p < .001, \text{Cohen's } d = 1.61) \) accuracy. Moreover, the direct group made significantly fewer grammatical \( (p < .001, \text{Cohen's } d = 1.85) \) and non-grammatical \( (p < .001, \text{Cohen's } d = 2.35) \) errors than pupils who practiced their writing skills during the treatment session (S2). Indirect CF also proved to have a more beneficial effect on both accuracy measures than either self-correction (grammatical accuracy: \( p < .001, \text{Cohen's } d = 1.05 \); non-grammatical accuracy: \( p < .001, \text{Cohen's } d = 1.33 \)) or writing practice (grammatical accuracy: \( p < .001, \text{Cohen's } d = 1.37 \); non-grammatical accuracy: \( p < .001, \text{Cohen's } d = 2.06 \)). The effects associated with direct CF were again greater than those of indirect CF. Analyses furthermore revealed that the interaction we found on our measure of overall accuracy – between the effectiveness of the different control treatments and pupils’ educational level – was related to non-grammatical accuracy only. The self-correction and practice treatments were equally (in)effective for pupils of both educational levels with respect to grammatical accuracy. However, the influence of the two control treatments on non-grammatical accuracy differed significantly across educational levels. Whereas the practice and self-correction treatments were equally effective for the higher-level pupils, the lower-level
learners benefited more from self-correcting their non-grammatical errors than from writing practice ($p < .001$, Cohen’s $d = 1.15$).

**Short-term learning effects (S3)**

ANCOVA’s revealed that there was a significant difference between groups one week after the treatment session (i.e. during S3) in performance on both the grammatical ($F (3, 231) = 4.10$, $p = .007$, $\eta^2_p = .05$) and non-grammatical ($F (3, 232) = 8.95$, $p < .001$, $\eta^2_p = .10$) accuracy measures. Post-hoc pair wise comparisons showed, however, that the effectiveness of direct and indirect CF differed across accuracy types (cf. Table 4.8). Both CF approaches were more effective in improving pupils’ non-grammatical accuracy than self-correction (direct CF: $p = .021$, Cohen’s $d = 0.53$; indirect CF: $p = .002$, Cohen’s $d = 0.68$) or writing practice (direct CF: $p = .003$, Cohen’s $d = 0.65$; indirect CF: $p < .001$, Cohen’s $d = 0.80$). In contrast, only direct CF helped pupils to reduce the number of grammatical errors in a new text which was written one week after the feedback had been provided; direct CF proved to be significantly more beneficial than practicing writing ($p = .016$, Cohen’s $d = 0.55$).

**Long-term learning effects (S4)**

When we considered pupils’ grammatical and non-grammatical performance on the delayed post-test (S4), we still found a significant between-group difference on both accuracy measures (grammatical: $F (3, 212) = 4.71$, $p = .003$, $\eta^2_p = .06$, non-grammatical: $F (3, 212) = 6.60$, $p < .001$, $\eta^2_p = .08$). Post-hoc pair wise comparisons revealed that the positive effect of direct CF on grammatical accuracy was still present four weeks after the treatment had taken place (cf. Table 4.8); pupils in the direct group made fewer grammatical errors than classmates who were allocated additional practice opportunity ($p = .004$, Cohen’s $d = 0.65$). Furthermore, whereas pupils’ non-grammatical accuracy benefited from both direct and indirect CF during the first post-test (S3), only the effect of indirect CF was sustained in the delayed post-test. Results showed indirect CF to be significantly more effective in reducing the number of non-grammatical errors in pupils’ writing than either self-correction ($p = .003$, Cohen’s $d = 0.66$) or writing practice ($p < .001$, Cohen’s $d = 0.82$).

**4.6.5 Effects of CF on written complexity**

We did not only aim at investigating the effect of CF on written accuracy, but also explored the influence of error correction on the structural complexity and lexical diversity of pupils’ writing. In doing so, we tested Truscott’s (2007) hypothesis that CF might lead to avoidance
of more complex structures. The initial ANCOVA models reported in this section contained treatment and educational level as between-subject variables, and language proficiency and either lexical or structural complexity S1 as covariates. Educational level and language proficiency were excluded from the final model since these factors turned out to be insignificant.

Our data did not confirm Truscott’s assumption that CF leads to simplified writing; we did not find any significant between-group differences on our measures of lexical diversity or structural complexity in either of the post-tests (structural complexity S3: $F(3, 244) = 1.45, p = .229, \eta^2_p = .02$; lexical diversity S3: $F(3, 244) < 1, p = .712, \eta^2_p = .01$; structural complexity S4: $F(3, 227) = 1.31, p = .271, \eta^2_p = .02$; lexical diversity S4: $F(3, 227) = 1.85, p = .138, \eta^2_p = .02$). Results did reveal significant between-group differences concerning the structural complexity ($F(3, 262) = 7.94, p < .001, \eta^2_p = .08$) and lexical diversity ($F(3, 262) = 31.65, p < .001, \eta^2_p = .27$) of the output produced during the treatment session (S2). The writing of pupils who wrote a new text during S2 (i.e. the practice group) was structurally less complex than the writing of pupils who received direct CF ($p = .001$, Cohen’s $d = .68$) or indirect CF ($p < .001$, Cohen’s $d = .77$). Furthermore, all pupils who had the opportunity to revise their text, outperformed the practice group on the measure of lexical diversity (direct CF: $p < .001$, Cohen’s $d = 1.55$; indirect CF: $p < .001$, Cohen’s $d = 1.32$; self-correction: $p < .001$, Cohen’s $d = 1.11$).

### 4.7 General Discussion

Truscott (1996; 1999; 2004; 2007; 2009) has repeatedly argued that CF has no place in L2 classrooms because grammar correction would be more likely to hinder than to facilitate accuracy development. Our findings, however, clearly show that comprehensive CF is an effective means of improving learners’ accuracy over time, and do not support Truscott’s supposition that CF has detrimental side-effects.

#### 4.7.1 The language learning potential of comprehensive CF

The main aim of the present study was to investigate whether comprehensive or unfocused CF leads to improved accuracy in L2 writing. Just as earlier studies (e.g. Ashwell, 2000; Fathman & Whalley, 1990; Ferris, 1997; Ferris & Roberts, 2001), we found that comprehensive CF enables learners to enhance the linguistic correctness of a certain manuscript during revision (RQ 1). The important contribution our study makes to the
existing literature, however, relates to the effectiveness of comprehensive or unfocused CF in expanding pupils' accuracy in new texts (RQ 2). Whereas research already revealed that learners receiving focused CF are able to use the targeted form more accurately in new pieces of writing (e.g. Bitchener, 2008; Bitchener & Knoch, 2008; Bitchener & Knoch, 2009; Bitchener & Knoch, 2010a; Bitchener & Knoch, 2010b; Ellis et al., 2008; Sheen, 2007; Sheen, 2010b), our study is the first to show that unfocused CF leads to learning; we found that pupils whose errors were corrected comprehensively made fewer errors in new pieces of writing than learners who did not receive CF. The fact that receiving CF proved to be more beneficial than self-correction without any available feedback, furthermore shows that CF has an added value above revision as such (RQ 3). Importantly, the positive effects of comprehensive CF showed to be durable; accuracy gains were visible both in the post-test and the delayed post-test, which respectively took place one and four weeks after pupils received the corrections. (Cf. Table 4.8.) Our post-test results indicated medium effect sizes (on average) for the advantage of CF over practicing writing and self-correction. This finding concurs with those of meta-analyses investigating the effectiveness of CF (Li, 2010; Lyster & Saito, 2010; Russell & Spada, 2006), which consistently reported medium effects of error correction on immediate and delayed post-tests, in both oral and written settings, and across laboratory and classroom-based studies.

4.7.2 Interactions between feedback methodology and error type

Apart from exploring the overall effectiveness of comprehensive CF, the present study investigated the relative efficacy of direct and indirect correction methodologies (RQ 4), and the differential effects of CF on grammatical and non-grammatical error types (RQ 5). We found that different CF types have value for different types of errors; whereas only direct correction promoted grammatical accuracy improvement, pupils' non-grammatical accuracy benefited most from indirect CF (i.e. the effect of indirect CF was retained the longest). (Cf. Table 4.8.) These findings are in line with suggestions in the literature concerning the relative effectiveness of direct and indirect CF. The cognitively demanding indirect approach did foster long-term acquisition of those linguistic features pupils were able and confident to self-correct (using the error codes provided), that is their non-grammatical errors (Bitchener

10 Whereas Lyster and Saito (2010) and Russell and Spada (2006) reported large overall effect sizes for CF, both studies concluded that CF has a medium effect on L2 acquisition; Lyster and Saito inferred that the effects of CF could be considered medium by subtracting the effect found for the control groups from the overall large effect of CF. When reconsidering the methodological quality of the studies included in their meta-analysis, Russell and Spada noted that studies reporting reliable and valid measures yielded medium effect sizes.
& Knoch, 2008; Ferris, 1995). Conversely, although indirect CF did supply learners with sufficient information to resolve their more complex, grammatical errors during revision, the fact that they did not know if their own hypothesized corrections were accurate might have prevented them from internalizing the correct structures (cf. Chapter 3; Chandler, 2003). Only when the grammatical errors were corrected directly, pupils showed evidence of long-term learning of the corresponding target forms. Based on these findings, we conclude that both grammatical and non-grammatical errors are amenable to CF, but that they benefit from different types of correction. This leads us to suggest an alternative explanation for the apparent lack of a learning effect in Truscott and Hsu's (2008) study. The error categories targeted in their study were mainly grammatical in nature; orthographical and lexical errors were not corrected. Moreover, an indirect type of CF (i.e. underlining of errors) was used. Our results indicate, however, that only direct CF has the potential to yield long-term grammatical accuracy gains.

4.7.3 The potential detrimental side-effects of CF
We also aimed at testing Truscott's (1996; 2004; 2007) hypothesis that CF could be harmful to learners' accuracy development, because (i) CF diverts time and energy away from more productive aspects of writing instruction, such as additional writing practice (RQ 6), and (ii) CF has a negative influence on the complexity of L2 learners' writing, in that it makes them avoid structures that have been corrected before (RQ 7). Our findings oppose both of these claims. CF did not lead our participants to produce lexically or structurally less complex writing, and the texts written by pupils who received CF were more accurate than those of learners who were allowed an extra opportunity to practice their writing skills.

4.7.4 Educational level and CF efficacy
A final issue the present study explored, was the possible influence of participants’ educational level on feedback effectiveness (RQ 8). The main reason for exploring this issue was practical in nature. However, it is also theoretically plausible that educational level mediates the efficacy of indirect CF. We presumed the higher-level pupils to dispose over more meta-linguistic knowledge than their lower-level peers. The hypothesis that indirect CF may be more helpful for learners with higher levels of meta-linguistic awareness

---

11 The fact that pupils in the self-correction group were able to improve their non-grammatical but not their grammatical accuracy during revision (S2) also suggests that the grammatical errors are the more complex ones.
(e.g. Ferris, 2004; Hyland and Hyland, 2006), then led us to predict that the higher-level pupils might be better able to use indirect corrections. Although we did find that the higher-level pupils outperformed the lower-level learners on the different linguistic measures used in the study, we never found a significant interaction between the effectiveness of the CF treatments and learners’ educational level.

4.7.5 Limitations and directions for further research

Although findings of the present study constitute clear evidence in favor of written comprehensive CF, we should keep in mind that it is limited in several ways. To begin with, our participants only received feedback once. Even though the pupils in the indirect group received instruction on how to use and interpret the error codes, they might not have had enough opportunity to get used to this type of correction. It is therefore possible that the effect of indirect CF would have been greater if we had applied a more longitudinal design with more than one CF occasion. In the same line of reasoning, we would not argue that practicing writing does not have any value for accuracy development; it could well be that practicing writing does lead to improved accuracy when learners are allowed more than just one additional practice opportunity. However, our results do not indicate that error correction is a waste of time in the sense that practicing writing is more beneficial (Truscott, 1996; 2004). On the contrary, even a single CF treatment proved to have long-lasting positive effects. These effects could only be expected to be greater if learners are offered CF on additional occasions.

Secondly, the reason that we did not find a significant interaction between pupils’ educational level and CF efficacy, might be that the difference between the levels included in this study was not big enough. It could be that a comparison of more dissimilar groups would have led us to other conclusions concerning the influence of educational background on feedback benefits. It also needs to be noted that the present study was performed within a particular context, which could be described as a relatively naturalistic SLA environment; pupils started learning Dutch at an early age, and the L2 was the means rather than the goal of instruction. As a result, our learners’ level of meta-linguistic awareness might have been too low to really and fully benefit from indirect CF, even for the higher-level pupils.

Thirdly, even though the present investigation was performed in intact classes, it could not be considered a real classroom study. The administered tasks – although representative of the written assignments used within the educational setting under investigation – were not part of the curriculum, and the feedback and task instructions were provided by the researcher instead of by the teacher. Consequently, we cannot be sure that
our findings would hold in a real-world class situation. On the other hand, we would not have been able to unambiguously establish the effectiveness of comprehensive CF in a less controlled setting.

Moreover, one could question our operationalization of grammatical accuracy. We included article errors, inflectional errors, word order errors, omissions, additions, and pronominal errors in our broad measure of grammatical accuracy. The syntactic errors concerning word order, and additions or omissions of constituents are undoubtedly problems that involve integral parts of a complex system, and therefore certainly belong to the type of grammar errors Truscott claimed (2007) could never benefit from CF. In comparison, errors in the use of articles, pronominals, and inflectional morphology might be considered less complex problems in relatively discrete items. However, analyses on a narrow category of syntactic errors only (i.e. word order errors, additions of non-necessary elements, and omissions of necessary elements) rendered the exact same results as the analyses on the broader measure of grammatical accuracy reported in this article. From this we conclude that comprehensive CF is effective in decreasing the number of even the most complex types of grammatical errors. Further research is warranted, however, to identify the exact effect of comprehensive CF on separate types of grammar problems, or even different functional uses of a single grammatical feature.

Finally, as Bruton (2009a) argued, comparing the error rates of two texts might not be the best way to investigate if and how learners have benefited from CF. In doing so one “assumes that the experimental group students could potentially draw on language knowledge in the second writing task gained from the feedback on the first writing task” (Bruton, 2009a, p. 137). It is uncertain, however, if learners can indeed use the knowledge they have gained from corrections on a particular piece of writing in a new text, on a new topic, in a new genre, and so on. Although our tasks were designed to be as similar as possible, it is still quite feasible that participants received corrections on features that they were unable to reuse when writing a new text. More detailed, qualitative analyses of learners’ accuracy performance over time might give a better picture of the accuracy gains brought about by CF. Research that explores this potential added value of an in-depth accuracy analysis would be worthy of further pursuit (cf. Chapter 5).

4.7.6 Conclusion

Despite the limitations listed above, our results clearly showed comprehensive CF to be effective in promoting both grammatical and non-grammatical accuracy during revision as well as in new pieces of writing, irrespective of learners’ educational level. Moreover, our
results did not constitute evidence of learners avoiding complex structures due to CF, nor did we find writing practice to be more beneficial than CF. Hence, the present study does not support Truscott’s (1996; 1999; 2004; 2007; 2009) claim that written CF is ineffective or even harmful. We conclude that comprehensive CF is a useful instrument that L2 teachers can use to help learners improve their accuracy in writing.