Computerised Speaking Practice: The Role of Automatic Corrective Feedback in Learning L2 Grammar
Computerised Speaking Practice: 
The Role of Automatic Corrective Feedback 
in Learning L2 Grammar

Proefschrift

ter verkrijging van de graad van doctor 
aan de Radboud Universiteit Nijmegen 
op gezag van de rector magnificus prof. dr. Th. L. M. Engelen, 
volgens besluit van het college van decanen 
in het openbaar te verdedigen op vrijdag 19 juni 2015 
on 10.30 uur precies

door

Bart Penning de Vries 
geboren op 3 oktober 1981 
te Bennekom
Promotor:
Prof. dr. R. van Hout

Copromotoren:
Dr. C. Cucchiarini
Dr. H. Strik

Manuscriptcommissie:
Prof. dr. M. van Mulken (voorzitter)
Prof. dr. A. Housen (Vrije Universiteit Brussel, België)
Prof. dr. J. Colpaert (Universiteit Antwerpen, België)

This work is part of the research programme Feedback and the acquisition of syntax in oral proficiency (FASOP), which is funded by the Netherlands Organization for Scientific Research (NWO).
And what is good, Phaedrus,
And what is not good --
Need we ask anyone to tell us these things?

- Robert M. Pirsig, _Zen and the Art of Motorcycle Maintenance_

[A]t the end of every commercial break, the show's trademark shot of planet earth as seen from space, turning, would appear, and the CBS day-time network announcer's voice would say 'you're watching As the World Turns,' which he seemed, on this particular day, to say more and more pointedly each time - 'You're watching As the World Turns' - until I was suddenly struck by the bare reality of the statement. I don't mean any sort of humanities-type ironic metaphor, but the literal thing he was saying, the simple surface level. I don't know how many times I'd heard this that year while sitting around watching As the World Turns, but I suddenly realized that the announcer was actually saying over and over what I was literally doing. Not only this, but I also realized that I had been told this fact countless times - as I said, the announcer’s statement followed every commercial break after each segment of the show - without ever being even slightly aware of the literal reality of what I was doing.

- David Foster Wallace, _The Pale King_
Acknowledgements

Working on the FASOP project towards this dissertation has been a great experience. Especially the experiments were exciting times, because that’s where the theory would meet the road, so to say. But I’ve also immensely enjoyed working with our FASOP group, which in turn was embedded in the friendly professional atmosphere of CLST.

First of all I’d like to thank everyone who participated in the experiments. Over two hundred people (thirty-five different nationalities) practiced with a version of GREET. It was always a lot of fun to meet everyone and receive your feedback. Your comments always sparked many ideas for improvements or for wholly new experiments.¹

For the recruitment of the learners, we depended heavily on the assistance of language institutes and schools. We are most indebted to Radboud In’to languages, whose classes we have visited on several occasions. A special thanks to all the teachers, and in particular to José Bakx for allowing our experiments, and to Petra Roël who reviewed the GREET language content. ROC Nijmegen was also very friendly and allowed us to run the experiment twice at their location, and especially teacher Pleuntje Gouwerok was super helpful. At Capabel Utrecht we’d like to thank Tjitske Zondag and Eefke Jacobs for helping us run the experiment with their students and booking the computer room. We ran a pilot experiment with low educated learners at Cumulus Welzijn Utrecht (de Bram). Here I’d like to thank Nienke Witkamp-Bruijn who took the time to send multiple emails to their ICT support to make sure the experiment would work. At Intercity-Nijmegen Henk van Ooij was great to work with as he invested much time to discuss with us the possibilities for an experiment, after which he recruited participants and set us up in their computer room. Thanks also to the James Boswell Institute, Utrecht, and especially Jolanda Straver for approving the experiment and Ine Kievits for recruitment in her class; and finally also the Language centre at HAN, where Manja Mekking ok’d the experiment and Eric Lemmens whose class we were allowed to visit.

Then secondly I’d like to thank very much the manuscript committee for carefully reading my dissertation over the Christmas period, and for their comments on the text.

For this project I was lucky to have very dedicated supervisors. Roeland, Catia and Helmer were virtually always available for advice or to promptly return useful feedback on texts. Roeland was always willing to discuss analyses and look for further explanations in the data, which was tremendously helpful and motivating.

¹ Maybe, if you’re a former participant and you’ve read chapters 3, 4, or 5, you might have found out that the participants got paid, while you were not! This means you were part of the first of two pilot experiments. We are immensely grateful to you and your participation improved the learning experience of all participants after you. But please contact me if you feel you are entitled to more than thanks.
VIII | The Role of Automatic CF in Learning L2 Grammar

Catia was great helping me out each time to structure my texts and activities. Helmer was always aware of the big picture and would make sure I stayed on the (straight and narrow) planned path. We were able to have many group meetings where we would discuss the project in detail. I think this resulted in a language learning experiment that covered almost all the bases and where it couldn’t, the best compromise was selected. This way Steve and I could work together both using experiments with GREET for our separate PhDs. Importantly, the atmosphere in our group was always relaxed but one of critical constructive thinking. There was much room for creativity and development of new ideas. Especially for designing the GREET program this has resulted in an experiment program that goes beyond what other CALL systems attempt to do. Way beyond. It just doesn’t make coffee.

Thanks to Joost van Doremalen who developed the speech recognizer and the first version of GREET. You’ve pioneered including Prince in a presentation, and you were also always good to talk to for a no-nonsense perspective on issues in language learning. Or on recording music.

Most of the development and experiment activities described in this dissertation were completed together with Steve. We worked together closely to develop a GREET system that could incorporate both our experiment plans. This has been a very enjoyable and fruitful collaboration and it extended well beyond simple PhD work into other important parts of life such as climbing, kick boxing, donuts, IPA beer, Twin Peaks (Lynch), On the Road, Lila, Cryptonomicon and Neuromancer. I think we also did some ground-breaking work on the key questions in horse-drawn CALL. Besides that important stuff, you were also great to work with. We had many useful meetings to hash out experiment issues (log output, motivation, planning) where the use of either flashcards or the white board was a must. Also the enthusiasm you brought to developing GREET was inspiring.

Our project also benefitted from work by our interns Judith, Polina, Kimber, Irene and Kimberly. Obviously so in their help with annotating (with special mention for winner of the prestigious ‘Best Annotator’-award Polina), but also in their projects using the GREET system to pursue various research questions.

Ineke van de Craats provided valuable advice when we were setting up the GREET experiments with low educated learners and in a later stage also helped to develop content for our longitudinal experiment.

I would also like to thank the two Marijkes (Bakkeren and Bleeker) at the MMS for helping us out with experiment logistics and allowing us all sorts of arrangements that weren’t traditional.

Working at the Radboud University was great because of the wonderful colleagues. The open atmosphere at the department ensured that it was easy to ask for and receive help on any research related issue. Special thanks to Henk, Hella, Vanja, Sara, Polina, Remy, Job, Mario, Christina, Eva, Michele and everyone else on the eighth and ninth floor of the Erasmus building.

For kick boxing workouts thanks to Steve, Yang and Eric. It seems that kick boxing and research go very well together. Especially when alternative hypotheses are needed for explaining the data. For transportation from the train station to the
Erasmus building, Eric generously provided me with a bike that I’ve semi-literally rode to rusty bits.

Perhaps as an aside, it is good to explain that GREET can be pronounced as ‘xre:t’, ‘gre:t’, ‘grɪ:t’, but not as ‘xri:t’. Based on the axioma that Dutch names sound cooler in American English (e.g. Van Halen) we initially chose the name GeRT (Grammatica en Regel Trainer), but due to the unintended phonological resemblance with a politician’s name that did little to enthuse Dutch learners, we opted for GREET instead. A Dutch name with a solid traditional feel to it with the added bonus of the characteristic velar fricative (\x), which serves well as a symbol of the challenges that learning L2 Dutch poses.²

During our experiments, we’ve had many types of problems, all of which we controlled for in the next, only for a new problem to arise. We’ve had many disruptions which are sort of amusing now, so for them to have some sort of use I’ll list them here so they won’t have been for nothing. Besides your more typical issues such as, computer crashed, server crashed, host server maintenance at the exact time of an experiment, server space full, local network too slow, automatic updates of all sorts of programs when it’s least wanted, flash incompatibility and security settings, port configurations and microphone malfunctions, we had the more intriguing problem that we labelled the matrix: speech recordings that were distorted into tinny and slurred sounds with a helicopter in the background. Still nobody knows what the matrix is. When all these issues were solved or pre-empted, we experienced a power out in the building, and in the experiment following, a fire drill.

Getting back on the thanks-track here.

For agreeing to be my Paranymp: Eric Sanders who is always the calm collected voice of reason in CLST except when going karaoke-ing in Tokio when Motörhead came on, and Marloes Gutenstein-Penning de Vries who continues to set examples for scientific achievement which I’m sure I’ll never reach, but in the spirit of competition will keep trying.

On a personal note, I’d like to thank my family and friends: basketballers (who are slowly turning into cyclists), the culinarios of the CZGZ, the intellectuals in the Boekenclub who push me to read Dutch lit., both bands for musical relaxation, and the guys who operate outside of these groups. In particular for advice on Dutch, Edith van Bodegraven, and Mirjam Streefkerk.

Finally my parents of course, who have always been supportive and gave me my copy of The Language Instinct long ago, and Shakespeare’s tragedies - and my drumkit. I’m sure that shows that they are awesome. The best for last: for loving support thanks to Quirijn who is never anything other than perfect and has been of great help during every step of the Phd enterprise.

² We cannot discount GREET’s part in the pleasure of FASOP, but we have to remember that it is a construct. It’s an it, not a she, as Dixie Flatline (Neuromancer, 1984) will point out to you.
X | The Role of Automatic CF in Learning L2 Grammar
Outline

Acknowledgements VII
Outline XI
Chapter 1: Introduction 15
  Introduction overview 15
  1. Corrective feedback: overview of research 16
  2. Corrective feedback and individual differences 20
  3. Corrective feedback and learner behaviour 22
  4. CALL environment for CF research 23
     4.1. Experimental environment 25
  5. Goals and outline of this thesis 30
Chapter 2: Adaptive Corrective Feedback in Second Language Learning 33
  1. Introduction 34
  2. CF in Second Language Learning 35
     2.1. Types of corrective feedback 35
     2.2. Explicit and implicit knowledge and learning 36
     2.3. Interface 37
     2.4. Noticing 37
     2.5. Learner differences 38
  3. Research on corrective feedback 38
     3.1. Overview of CF research 39
     3.2. Effects of different types of corrective feedback 39
     3.3. CF and individual differences 40
  4. New opportunities for implementing and investigating CF in L2 learning 41
     4.1. CALL systems 41
     4.2. An ASR-based CALL system for studying the effect of CF on oral proficiency 42
  5. Conclusions 45
Chapter 3: Spoken Grammar Practice and Feedback in an ASR-based CALL System 47
  1. Introduction 48
  2. Research background 48
     2.1. Spoken grammar practice through computer assisted language learning 49
     2.2. Language production and second language learning 49
     2.3. Research questions 51
  3. Method 52
     3.1. The GREET system architecture 52
     3.2. Participants 53
     3.3. Procedure 54
     3.4. Target structure 54
     3.5. Treatment design 55
     3.6. Proficiency tests 57
Outline | XIII

2.3. The FASOP (GREET) system 112
2.4. Treatment 113
2.5. Proficiency test: discourse completion task 114
2.6. Learner evaluation questionnaire 114
2.7. Logged data of learner behaviour 115
3. Results 115
3.1. Participant background 115
3.2. Discourse completion task (DCT) 116
3.3. Learner evaluation 119
3.4. Learner behaviour 120
4. Discussion and conclusion 124

Chapter 6: Discussion and Conclusions 129
1. Introduction 129
2. Prerequisite methodological innovations 129
3. The role of CF in learning oral L2 grammar 132
5. What is the impact of CF on learner behaviour? 136
6. General conclusions 138
7. Limitations 139
8. Future research 142

References 145

Appendix 161

Appendix A: Grammaticality judgment task items (Chapter 3) 162
Appendix B: Discourse completion task items (Chapter 3) 164
Appendix C: GREET treatment target Items (Chapter 3) 166
Appendix D: Post-test questionnaire (Chapter 3) 168
Appendix E: Grammaticality judgment task items (Chapter 4 & 5) 169
Appendix F: Discourse completion task items (Chapter 4 & 5) 172
Appendix G: Pre-test questionnaire 175
Appendix H: Post-test questionnaire items (Chapter 4 & 5) 176
Appendix J: Treatment Items (Chapter 4 & 5) 178

Nederlandse samenvatting 185

Inleiding 185
Correctieve feedback 186
GREET 187
Overzicht van het proefschrift 188

Curriculum Vitae 193
Chapter 1: Introduction

Introduction overview

The role of corrective feedback (CF) in second language (L2) learning has been a source of discussion for many years. In particular, the contribution of CF to improving oral proficiency has been a topical issue. Several researchers have claimed that providing CF does not contribute to improving language proficiency at all (Krashen 1982; Truscott 1996,1999; Schwartz 1993; VanPatten 1989) putting the burden of proof on theorists that claimed CF was effective (e.g. Long 1996; Gass 1997; Pica 1994; DeKeyser 1998). Theories also disagree on the characteristics of CF, or processes that CF invokes, that are beneficial for L2 learning (e.g. Lyster & Ranta 2013 versus Goo & Mackey 2013). As a result, empirical research on the effect of CF on oral L2 proficiency is necessary to test competing claims in SLA theory. Besides its relevance for SLA theory, CF research may have clear pedagogical implications.

Questions on the role of CF are of continuing interest, as empirical studies on oral CF have provided varying outcomes. Meta-analyses of studies on oral CF reported differences in, e.g., context (classroom, lab, face to face, and more recently computer environments), method (operationalization of CF, type of treatment, measurement of proficiency), and indicated the importance of so far less investigated variables such as individual differences. As a result, the meta-analyses have been tentative in their suggestions that CF is effective, and more tentative still on the variables that contribute to this effectiveness.

Several researchers, including the authors of the meta-analyses, have therefore called for further research on CF under more controlled conditions, to gain more detailed knowledge on how each of the variables mediates CF effectiveness (e.g. Hulstijn 1997; Russell & Spada 2006; Li 2010; Goo & Mackey 2013). More attention should be paid to learner differences (Dörnyei 2005; Long 2007; Li 2010) and to the language learning process (e.g. how a learner interacts with CF ) (Mackey & Gass 2006). These requirements can be addressed by resorting to a computer assisted language learning (CALL) system, where learners practice individually with controlled input and automatic feedback and where they can be monitored during practice (Heift & Schulze 2007; Presson, Davy & MacWhinney 2013; Chapelle 2009b). However, in order for the CALL system to provide automatic CF on oral production, it has to be equipped with automatic speech recognition (ASR). There are clear benefits to employing ASR for CF research, but so far there are very few studies that make use of ASR to provide automatic CF (Bodnar, Cucchiarini & Strik 2011; Golonka, Bowles, Frank, Richardson & Freynik 2013; Clifford & Granoien 2008).

In this thesis, we respond to the call in the SLA literature for more controlled research on the role of CF in the acquisition of oral proficiency. We developed a CALL system that makes use of ASR to provide automatic CF on oral production
and that records all user-system interactions. As research shows that the target linguistic feature may influence CF effects, we focus on a specific grammatical feature of Dutch grammar, namely verb second (V2).

In this introductory chapter we explain how this thesis addresses outstanding issues in CF research. We have three focus points, briefly introduced here in Chapter 1. In Section 1 we explicate the need for controlled environments for research on the effectiveness of CF on oral proficiency. Section 2 addresses the relevance of individual differences in CF research, while Section 3 shows the need for CF research that investigates the learning process in addition to the learning results. In Section 4 we discuss how these needs can be met in a CALL system, and we then briefly introduce the ASR-based CALL system that we developed and used in the research reported on in this thesis, GREET. Section 5 presents the goals of this thesis, followed by an overview of the chapters in Section 6.

1. Corrective feedback: overview of research

In this section, we will briefly provide an overview of the issues surrounding CF research and how they are addressed in our research. We find that (uncontrolled) variability in CF research is an important issue, and that there is a need for controlled research environments.

Definition and role of CF

Corrective feedback (CF) is information given to learners regarding a linguistic error they have made (cf. Loewen 2012; R. Ellis 2006). This information can come from different sources, such as an interlocutor or a teacher; in person, or through a CALL system. The source of CF, as well as the context in which it is given, have been found to influence the way CF is received (Sheen 2004, 2008).

There are various ways in which CF can be provided. Lyster and Ranta (1997) analysed oral CF given in an L2 immersion classroom and categorised it into six types. This taxonomy was applied in numerous subsequent studies, but since the six CF types had not been rigorously defined, interpretations and operationalizations of CF types turned out to vary across studies (e.g. Nicholas, Spada, Lightbown & 2000) discuss the differences in recasts across studies). In a broad sense, CF types are classified along two dimensions: a) the degree of implicitness or explicitness, and b) whether they provide input or prompt output (see Lyster, Saito & Sato 2013). The differences in types of CF influence the learners’ reaction to CF, and its effectiveness (Mackey & Goo 2007). To illustrate the characteristics of CF four examples are presented. In example (1), the teacher reformulates the learner’s error (provides input), thus implicitly signalling the presence and location of the error. In (2) the teacher indicates the presence of an error, and elicits a correction. In (3) the teacher explicitly indicates the error and provides the correction. In (4), which is an example from our CALL system, the learner is explicitly told there is an error, and prompted to correct the error.
Different theories in SLA ascribe effectiveness of CF to different characteristics. In classrooms, implicit reformulation feedback is most frequently used by teachers (Lyster & Ranta 1997). This is because it minimally impacts the flow of communication, but at the same time it is likely to be ignored or misinterpreted by the learner; explicit feedback is less likely to be misinterpreted, but is more disruptive for communication and more likely to cause anxiety in the learner. L2 theories based on skill-acquisition theory predict that CF that prompts output is most beneficial (cf. uptake, Lyster 2004; DeKeyser 2003, 2007) whereas interactionist theory stresses the importance of modified input during negotiation for meaning as a result of CF (Long 1996; Gass 1997; Pica 1994). Sociocultural theory posits that different types of CF are necessary at different stages of acquisition (Aljaafreh & Lantolf 1994). So far, these issues are unresolved, as findings on the effectiveness of CF type vary across studies (see for an overview R. Ellis, Loewen & Erlam 2006).

Besides the effectiveness of CF type, learners and teachers have different beliefs and preferences for providing and receiving CF (see Schulz 1996; Jean & Simard 2011; Truscott 1999; Lyster et al. 2013). How learners perceive CF is an important factor in how they experience L2 practice (Bodnar, Cucchiariini, Strik & van Hout 2014) which subsequently may consciously or subconsciously influence the effectiveness of L2 practice (Kartchava 2012).


**CF and Noticing**

Among L2 learners, it is generally assumed that immersion in the target language environment is the easiest way to learn an L2: that, by relying on the same mechanism as for learning the first language (L1), the L2 automatically emerges. Though there is no contention in the SLA literature that the L2 is for a large part acquired through exposure (cf. N. Ellis 2006), there are crucial differences with L1 acquisition. Whereas this latter process leads to uniform success in the L1, adult learning of L2 is characterized by variable levels of ultimate attainment (Birdsong 1992). Even with sufficient exposure to L2 input, learners usually do not fully acquire grammatical accuracy (Swain 1985; N. Ellis & Larsen-Freeman 2006).

Particular features of the L2 may not be acquired for several reasons, for instance, due to interference of the L1, low salience of grammatical forms, the less than perfect contingency between their forms and functions, etc. (N. Ellis 2006; N. Ellis & Larsen-Freeman 2006). This is formulated in the Noticing Hypothesis (Schmidt 1990, 2001), which says that “input does not become intake for language learning unless it is noticed, that is, consciously registered” (Schmidt 2001, p. 3).

To improve L2 learning, both in terms of efficiency and higher levels of ultimate attainment, a necessary step seems to be to facilitate noticing (Skehan 2003). This means that the learner’s attention occasionally needs to be directed to language form (cf. Long & Robinson 1998; Long 1990; Skehan 2003). An important method of facilitating noticing of features is through CF.

By drawing attention to language form (making the learner consciously notice an error), CF should result in explicit learning. Implicit and explicit learning are said to lead to distinct types of knowledge: implicit knowledge, and explicit knowledge (N. Ellis 2005). Implicit knowledge allows for rapid access and underlies language (spoken) language proficiency, and explicit knowledge is consciously accessed and only available for offline production. For CF to be effective for L2 (spoken) proficiency, it should have an effect on implicit knowledge.

Researchers who argued that CF was irrelevant, and even detrimental to L2 learning (Krashen 1982; Truscott 1996, 1999) stated that L2 knowledge that was not learned through implicit (subconscious) learning could not result in implicit knowledge (i.e. the non-interface position). However, studies suggest that explicit knowledge can support the development of implicit knowledge (DeKeyser 1998; Hulstijn 1995), but how CF influences implicit knowledge is still a matter of debate (i.e. through automatization (DeKeyser 1998, 2007; De Bot 1996), or through increased saliency in the input as a result of frequency or explicit knowledge (N. Ellis 2002; VanPatten 2004)). In addition, the way in which implicit knowledge should be measured is a topic of discussion as well (R. Ellis 2005; De Graaff & Housen 2009; Spada & Tomita 2010; Rebuschat 2013).

---

3 The role, and necessary degree, of conscious involvement for noticing is a subject of discussion, cf. Tomlin & Villa (1994)
Effectiveness of CF

As a result of the theoretical discussion on CF and its pedagogical significance, many studies on CF have been conducted. These studies varied in their design, level of control, and statistical power. In response, several meta-analyses attempted to synthesise overall findings.

A first meta-analysis was conducted by Norris & Ortega (2000), who examined 49 studies on language instruction, and within these studies they analysed CF as instructional treatment. They found that CF was effective, and that explicit (metalinguistic) CF was most effective. However, in the discussion they point out that the methods and practices in the included studies were widely variable "and generally not conducive to the systematic accumulation of knowledge about particular variables" (p. 499). Russell and Spada (2006) conducted a meta-analysis of 15 studies (laboratory and classroom, excluded CALL studies). They found an overall effect of CF, stronger for written than for oral CF. However, they observe that “the wide range of variables examined in CF research is spread rather thin; more work is needed to consolidate efforts and focus on those CF variables that appear to be particularly fruitful for future investigation” (p. 156). In an effort to expand the findings by Russell and Spada (2006), Mackey and Goo (2007) conducted a meta-analysis of 16 studies on CF (face-to-face, and two computer mediated communication studies) and included the variable CF type. They found an overall effect of CF on the short term (no effect on the delayed post-tests), and suggested that implicit CF (recast) is more effective than explicit CF (metalinguistic and explicit correction). Li (2010), in a meta-analysis of 33 studies (face-to-face, and computer delivered CF) found a medium overall effect of CF, and the effect of implicit feedback was better maintained than that of explicit feedback. Finally, Lyster and Saito (2010) present a meta-analysis of 15 classroom-based studies. They found that CF (explicit correction, recasts, prompts) was overall more effective than no CF, and that prompts were most beneficial.

Several issues are raised by the authors in their discussions of their findings. Due to small numbers of studies on each of the variables, authors caution for generalizability. Variables that were found to mediate CF effects between studies are the following ones: research context, research setting, task type, treatment length, interlocutor type, CF types, learner age (Russell & Spada 2006; Lyster & Saito 2010; Li 2010). Researchers operationalized feedback differently (Li 2010), which makes effectiveness of types across studies difficult to compare. Norris and Ortega (2000) criticised the studies in their meta-analysis for making claims about implicit knowledge, while they used outcome measures that could be argued to measure mostly explicit knowledge. In a follow-up on this issue, Doughty (2003) further argued that until studies include more measures of implicit knowledge, we cannot be confident that instruction leads to L2 competence that is unconscious, unanalysed,

---

4 This suggestion is questioned by Spada & Tomita (2010), since this claim is based on only a few studies.
and available for use in rapid, spontaneous communication. According to Spada and Tomita (2010), more studies currently include measures of implicit knowledge, though there is still discussion on which tests are most reliable for measuring implicit knowledge.

These issues have resulted in a (repeated) call for more controlled studies (e.g. Hulstijn 1997; Russell & Spada 2006; Goo & Mackey 2013), that can more reliably build on each other, preferably through replications of studies. Li (2010) calls for research on "learners of languages other than English, ... that is implemented in the computer mode, ... and the variables that moderate the effects of corrective feedback on SLA, such as age, gender, proficiency, L1 transfer, culture, complexity of the target structure, or interlocutor type, etc." (p. 349). Through rigorous experimental designs that control the research conditions single variables can be reliably investigated.

Clearly, however, language learners are distinct because of many relevant individual differences. These individual differences cannot be controlled, but should be accounted for, as they interact with the language learning environment (Dörnyei 2005).

2. Corrective feedback and individual differences

The variables that interact with CF to determine its effectiveness can be distinguished in learner internal and learner external variables. The meta-analyses show that external differences influence the results. However, individual differences, the learner internal variables, also play an important role in L2 learning; they have been found to be consistent predictors of L2 learning success (Dörnyei 2005). According to R. Ellis (2010) “The vast bulk of CF studies has ignored learner factors, focusing instead on the relationship and the effect of specific CF strategies and learning outcomes. … The general neglect of individual difference factors in CF studies is surprising because it would seem self-evident that learners will vary considerably in how they respond to CF, whether oral or written.” (p. 339).

Studies that have investigated individual differences found that they interact with feedback (e.g. learner proficiency (Mackey & Philp 1998; Mackey 1999), working memory (Mackey, Philp, Fuji, Egi & Tatsumi 2002), motivation (Uzum 2010), anxiety (Sheen 2008)). Providing learners with CF appropriate to their individual needs is likely to improve learning results (Cronbach & Snow 1977; R. Ellis 1994; Doughty & Long 2003). A successful example is given in Han (2001). The opposite can also apply that CF can be inappropriate for a particular learner, and not be effective at all, for example Lyster 2004 argues that recasts are ineffective in classroom setting.

So far it cannot be easily determined what type of feedback is most effective for a learner, not in the least because these needs may be constantly changing as proficiency develops (R. Ellis 1994). Research is needed to determine when CF - and what type of CF - is most effective for each type of learner at each stage of acquisition. In order to determine that, only learning outcome results may not be
sufficiently informative. Data on learner behaviour - how participants respond to and use CF - are required to provide additional information (this is discussed in Section 3).

The role of CF should therefore be examined in interaction with individual differences. In Chapter 4, we present a detailed study on CF effects, taking into account learner background and proficiency, and in Chapter 5 we look at the interaction of educational background with CF effectiveness.

**Proficiency level**

Depending on the proficiency level of the learner, the requirements for CF may change. Anmar and Spada (2006) found that learners at lower levels of proficiency benefited more from feedback techniques that explicitly signalled the nature and location of errors. This was also found to apply to learning of two different grammatical structures in Li (2013): low level learners benefitted most from metalinguistic feedback as compared to recasts, high level learners benefitted equally from both types of CF. In some cases, therefore, studies on (types of) CF may produce different results depending on the proficiency level of their participants.

Learners also notice features in the L2 without CF, for instance in the input or in their output. Also in those circumstances, learner proficiency plays an important role. More proficient learners are found to notice more of their errors in the output (Hanaoka 2007). For input in general, Carroll (1999) argued that intake is determined by our grammars and we "perceive what our linguistic systems enable us to perceive—based on the categories and structures of the grammar" (p. 343). As such, for language learning with and without CF, proficiency level influences L2 learning. This is important as learning may also take place in an experimental condition without CF.

Proficiency, however, is a broad concept and so far eludes a clear definition (see Housen & Kuiken 2009 for a discussion on proficiency as defined by the components of complexity, accuracy and fluency (CAF)). To address this problem for experimental purposes, proficiency level needs to be defined in some form of a pre-test (cf. Li 2010). In addition, proficiency should be measured on the basis of the target feature, as CF effect can differ according to the grammatical structure under study (R. Ellis 2007; Li 2013)

**Education level**

Language proficiency is important for socio-economic integration into the L2 community (Dustmann & van Soest 2002). Learners with a lower educational background, however, generally reach lower levels of L2 proficiency (Young-Scholten 2013). The literature shows that most CF research is conducted with high-educated learners. Consequently, the findings on CF effectiveness so far are based almost exclusively on high-educated learners (Young-scholten 2013). On account of
the individual differences found among high-educated learners with respect to L2 learning (e.g. metacognitive skills (Shute 2008) and working memory capacity (Mackey & Philp 1998) influence CF effect), the findings from SLA may not generalize to low educated learners. One of the few studies on low educated learners (Bigelow, DelMas, Hansen & Tarone 2006) showed that their noticing of the corrective intent of CF (recasts) was lower than that in a similar study with higher educated learners (Philp 2003).

In Chapter 5 we compare learning with and without CF for learners of different educational levels. By inspecting both learner behaviour and learning outcomes, we discuss the differences between learners of different educational levels.

3. Corrective feedback and learner behaviour

SLA perspectives such as sociocultural theory, interactionist theory, and dynamic systems theory emphasise the necessity of taking the learning process into consideration, besides learning outcomes (Aljaafreh & Lantolf 1994; Mackey & Gass 2006; De Bot, Lowie & Verspoor 2007). The effect of CF can be measured in pre- and posttest comparisons after an instructional treatment. In such an analysis, treatment is considered as a sort of a 'black box' which causes changes in learner proficiency. However, how learners behave during treatment, how they process and react to CF, can be an additional source of information to complement the outcome measures.

CF is a response to a learner error, which may in turn have an effect on the learner's language behaviour. This effect may be short term, for instance the repair of the error, or longer lasting by impacting the learner's interlanguage grammar, for instance by resulting into a lower frequency of the error. Learner behaviour can therefore provide information about the effectiveness of CF (Hegelheimer & Chapelle 2000).

Various processes that are assumed to be beneficial for L2 learning can be triggered by CF. CF facilitates the noticing of errors (Schmidt 1990, 2001), it can trigger output in the form of uptake (modified output, or repair) (Swain 1985; De Bot 1996; DeKeyser 2007), it can provide the learner with (modified) input (Gass 1997; Long 1996), and it can impact the learner’s affective state (Sheen 2008; Bodnar et al. 2014). Monitoring learner behaviour yields information about these processes. Individual differences are also found in interactions with the L2 practice environment (Dörnyei 2005; Heift & Rimrott 2012). Inspection of differences in learner behaviour can provide more detailed information on differences measured in learning outcome.

A controversial point in the CF literature is the role of uptake. Its importance was suggested by Lyster and Ranta (1997) in their analysis of teacher CF types in the classroom. Uptake is defined as “a student’s utterance that immediately follows the teacher’s feedback and that constitutes a reaction in some way to the teacher’s intention to draw attention to some aspect of the student’s initial utterance” (Lyster & Ranta 1997, p. 49), and they suggest that uptake may be a beneficial by-product
of CF. Repair of the error may suggest that the learner has noticed his/her error (cf. Hegelheimer & Chapelle 2000), but production of uptake itself can be beneficial. R. Ellis, Basturkmen and Loewen (2001) suggest that uptake may facilitate acquisition by "providing opportunities for learners to proceduralize target language knowledge already internalized in declarative form" (p. 282).

However, as Ammar and Spada (2006) observe, several researchers (including Lyster) “acknowledge that immediate repair is a questionable yardstick by which to judge learning because its absence cannot be taken as evidence of lack of learning” (p. 546). According to Gass (2003, p. 236) repair following recasts might be a sign of “mimicking”: repetitions that do not involve any analysis or revision of L2 knowledge. According to interactionist theory, the beneficial role of CF is to trigger negotiation of meaning and to provide modified input, and as such, uptake itself is not effective. Currently, the role of uptake is still unresolved (cf. Goo & Mackey 2013 versus Lyster & Ranta 2013). Studies on the effectiveness of uptake are few. According to Mackey and Philp (1998) uptake does not yield additional learning, whereas Loewen (2004) in an analysis of learner interaction finds that uptake can have an effect on later language use.

Consequently, uptake as a result of CF cannot directly be equated with effective learning behaviour, and more research is needed to understand its role. Analysis of learner behaviour can indicate the amount of received input, produced output, and instances of uptake (repair, modified output). By relating this information to learning outcome measures, inferences about their contribution to effective practice can be made.

4. CALL environment for CF research

As discussed above, research on CF can benefit from a controlled environment that takes individual differences into account, and which allows for analyses of learner behaviour. A CALL environment offers possibilities to meet these requirements.

---

**Figure 1. A componential framework for investigating CF** (adapted from R. Ellis (2010, p. 337) where the diagram applies to oral and written CF).
As seen in the previous sections, CF effectiveness is mediated by many variables. In Figure 1, a framework for CF research suggested by R. Ellis (2010), two groups of factors interact with CF to determine how a learner engages with the CF, which in turn determines the learning outcome. In order to understand how and when CF is effective, it is necessary that each of these factors be studied independently and in relation to each other. In a systematic approach, these factor groups should be controlled or monitored so that CF effectiveness can be assessed.

**CALL as a controlled environment for research on spoken language learning**

As will be discussed in more detail in Chapter 2, a CALL environment allows for much control over the contextual factors, and the learners interact individually with the system. In such a context, the input to the learner can be held constant, and the reactions by the system (CF) to the learner can be clearly operationalized and made consistent. In addition, the learner interactions can be logged. This way, the learning outcomes can be studied as the result of controlled practice, but also learner behaviour during practice can be inspected. Another benefit of a CALL experiment is that it is replicable. Taken together, a CALL system can be an optimal experiment setting (Hegelheimer & Chapelle 2000; Presson, Davy, MacWhinney 2013).

Within CALL there is a difference between the computer as a medium and as a tutor (Levy 1997). The computer can act as a medium, to allow learners to interact with each other (CMC), or as a tutor, where the learner interacts with the computer. In CMC, learners can engage in written or spoken chat in their L2, and the interactions and CF can be analysed in detail (e.g. Loewen 2004). In tutorial CALL, the learners interact individually with a virtual tutor that is programmed to react in a specific way. Though the interactions are more constrained, this has the benefit that the learner receives immediate and consistent CF, which is not the case in CMC (Dickenson, Eom, Kang, Lee & Sachs 2008). The consistency of practice is important for the comparison of individual differences in L2 learning. Given the model in Figure 1, we can investigate the individual difference factors by controlling the CF and the contextual factors and analysing how they engage with the CF.

Interestingly, "despite a vast interest in studying the role of [CF] in the oral classroom, very little research has been conducted in the CALL environment" (Heift & Schulze 2007, p. 154). Moreover, most CALL research addressed the written modality. Though there is evidence to suggest that written language practice can influence oral proficiency (Payne & Whitney 2002), CF on oral production is different from CF on written production, despite some overlap (Sheen 2010). In addition, spoken production is necessary for skill-specific practice (DeKeyser 2007), and for research on the role of uptake as a result of CF.

So far, it has proven difficult to conduct CALL studies on spoken language. To make this possible it is necessary to use speech technology. As seen in Presson et al. (2013, p. 153), there is still mistrust of the use of speech technology for L2 practice.
This feeling is misplaced and often caused by inappropriate use of ASR. As Holland & Fisher (2007) and Neri, Cucchiarini and Strik (2002) argue, ASR can be effective, if its limitations are taken into account in the design. Prior research has shown that ASR can also provide reliable results for low proficiency L2 speakers (Doremalen, Cucchiarini & Strik 2010).

**Learner behaviour in CALL**

A valuable contribution of CALL to SLA theory and pedagogical research lies in its capabilities of tracking learner behaviour (Fischer 2007; Collentine 2000; Hulstijn 2000; Heift & Schulze 2007; Chapelle 2007). Learner behaviour can be analysed to look for indications of when learning takes place (Collentine 2000; Mackey 2006; Chapelle 2007), and inferences about noticing can also be made through investigation of logged data in a CALL context (Hegelheimer & Chapelle 2000; Chapelle 2005; Presson, et al. 2013). Heift and Schulze (2007) distinguish between CALL studies that are performance based and those that are interaction based (p. 154-5). Where the former look at the effects of different types of feedback on the learning outcome, the latter are primarily concerned with the ways in which participants/ students use CALL applications. In Chapter 3, 4, and 5 we show that these two ways of analyses can be combined to triangulate the results, and have learning outcome data informed by learner behaviour, and vice versa.

Through analysis of learner behaviour, and relating behaviour to learning outcome, it is possible to identify which behaviour (and which CF interaction) is most conducive to L2 learning. This has both relevance for SLA theory (e.g. for uptake), and pedagogical relevance for the development of more efficient CALL systems.

**4.1. Experimental environment**

We identified the requirements from CF research, and determined that a CALL system that makes use of ASR can be an optimal experimental environment to study CF on spoken grammar. For our study, we developed such a system, which we named GREET. With this system, learners practiced speaking with and without CF. In addition, we developed proficiency tests and questionnaires that gauged the learners’ progress and appreciation. For a detailed description of the system, we refer to Chapter 3, but here we will provide a brief overview and a description of its development.

**ASR-based CALL system (GREET)**

The research reported on in this thesis was aimed at investigating the role of CF in L2 spoken learning. To provide controlled and consistent automatic CF on the
grammaticality of spoken learner output we employed ASR. At the beginning of our project in 2010 there were no systems offering spoken grammar practice and CF (Bodnar et al. 2011). Research at our department In Nijmegen had shown that it was possible to provide pedagogically valid (Neri, Cucchiarini, Strik, Boves 2002) and accurate CF to (low proficient) L2 learners with different L1 backgrounds (van Doremalen, Cucchiarini, Strik 2010; Strik, van Doremalen, Colpaert, Cucchiarini 2013), provided that the limitations of ASR technology were taken into account.

For pedagogical and experimental purposes, the demands on CF accuracy are high. Learners should not receive information that their response was incorrect when it was correct (False Reject), or that it was correct when it was incorrect (False Accept). As a result, it was not possible to allow learners free-response exercises. To ensure ASR accuracy, especially for learners of different L1 backgrounds, it was required that the number of possible utterances were constrained (providing automatic CF on free-response exercises would require accurate grammatical parsing of unpredicted utterances, which is currently insufficiently accurate). To enable accurate CF on spoken production, therefore, we designed exercises where the possibilities for output were restricted, while still requiring the learners to construct an utterance. In the exercises, learners were given segments of sentences, which they were required to assemble into a whole sentence. In this way, learners could practice Dutch word order and our target feature (verb second, see below), and the system could provide CF.

Taking into account the requirements for accurate ASR thus resulted in design requirements for the language content. We decided to use a virtual teacher-user interaction, as opposed to a virtual peer-user interaction, to make the CF that the learner received feel more plausible. From this starting point we developed a language lesson in the form of a dialogue. A major determiner of the success of language practice is the content of the lesson. Educational publisher Malmberg provided us with the language learning videos from their online course ‘Nieuwe Buren’. In cooperation with Dutch L2 teachers at Radboud In'to Languages, we developed question and answer dialogues that were based on clips from the video, and were appropriate for learners at the A2 CEF level.

For our language exercises, we initially wanted to select a CF type that was most likely to be effective. However, as seen in the literature overview, there is no clear answer: both recasts and prompts have been found to be most effective. We selected prompts on the basis of (1) their operationalizability in CALL, and (2) the similarity of the CF type as it is realized in the classroom. Recasts are automatically more explicit in a CALL setting, and thus would be clearly different from recasts in other contexts. The third argument is (3) their propensity to elicit modified output. In order to divert some of the learners’ attention from the fact that the CF only addressed language form, we included some multiple choice questions about the content of the video, to draw some attention to meaning.
Figure 2. A screenshot of the spoken language activity in GREET

The learner interacts with the system using the interface shown in Figure 2. The learner is asked questions about a film clip: 30 s clips of an ongoing story about a man moving to a new home in Amsterdam. Using ‘word blocks’ on the screen they form a spoken response, which they record. The recorded utterance is sent to the speech recognizer, which outputs a recognition result. This is mapped to a sentence in the language model that is tagged for grammaticality. The system then presents the appropriate feedback message to the learner. CF is provided as explicit notification of an error and a prompt to retry (see Example (4) above). In addition, to assist the learners to reach the correct answer, the system incrementally fills in a part of the sentence after each error.

The CALL system is setup in separate components (the speech recognition server; the courseware database; and the courseware engine that interacts with the client) so that it can easily be adapted to accommodate other types of experiments (though in this thesis it is kept constant). For instance with more, or other language content (videos and questions), linguistic target structures, and other types of feedback.

Finally, the system contains a logging module, which records all the learner’s interactions with the system. In a CALL system, the interactions can be logged in great detail. From the time the learners log in to the program (or website) until they log out, all interactions can be timed and recorded. This can be used for analysis of learner behaviour to inform L2 learning effectiveness and system performance (i.e.
is the system working as expected). In our study we recorded the learner’s responses (as sound files), the time they spent per page and time between certain actions, and the number of actions related to linguistic (target) items in the treatment. These were all available for analysis. In Chapters 3, 4, and 5 we discuss the data that have been logged.

In order to gauge the learners’ experience using GREET for L2 practice, the last task in all experiments was a post-test questionnaire. The questionnaire asked the participants about enjoyment and perceived effectiveness, and in open questions asked them for preferences and comments.

**Target feature**

R. Ellis (2007) shows that the effect of metalinguistic CF can vary for different grammatical structures (past tense -ed, comparative). For spoken interaction, Li (2013) found that the effect of CF interacted with the target structure and the learners’ proficiency level. As a result, the grammatical target is relevant for studies on CF effectiveness.

In all experiments reported in this thesis we investigated the same grammatical feature, verb-second (V2). As a result of V2, Dutch declarative sentences have a finite verb in second position. In Dutch, the subject precedes the finite verb, but when a constituent other than the subject is in the first position of a main clause, the order of subject and verb is inverted, as illustrated in (5), no inversion, and (6), with inversion.

\[
\begin{align*}
(5) \quad Bob & \text{ koopt een brood} \quad (\text{‘Bob buys a bread’}) \\
S & V \quad O \\
(6) \quad Morgen & \text{ koopt Bob een brood} \quad (\text{‘Tomorrow Bob buys a bread’}) \\
A & V \quad S \quad O
\end{align*}
\]

V2 is an appropriate feature to study, as the acquisition of subject-verb inversion presents a major problem for second language learners of Dutch (Jordens 1988). The reason seems to be that V2 does not affect the meaning of the sentence, and is therefore likely to be non-salient in the input and unlikely to prompt CF in meaning-based communication. Consistently redundant grammatical items take the longest for learners to acquire (R. Ellis 1994; Gass & Selinker 2008). A syntactic account suggests that V2-inverted order is more derivationally complex than V2-straight order in Dutch: it requires more movement (Blom & De Korte 2011). By default learners may opt for the less complex sentence, and will continue to use it until CF triggers noticing of the error.

With our system we can test learners with various L1s and see if the treatment interacts with L1 (or any other, additional language) background. The V2 target feature that we investigate is found in some L1s. The L1 is a factor that interacts with L2 acquisition (N. Ellis 2006; Rebuschat 2013), and knowledge can transfer from the L1 to the L2 (Odlin 1989; Ringbom 1992). This may be affected by the
learner’s knowledge of another L2, for example a German learner with L2 English (non-V2) may not automatically transfer the V2 feature to Dutch, on account of their knowledge of English (cf. Bohnacker 2006). It is important to take the learner’s language knowledge into account, therefore.

**Proficiency tests**

Some previous research on CF has been criticised for employing tests that were considered measures of explicit knowledge (R.Ellis 2007; Norris & Ortega 2000; Doughty 2003). It is important to determine the effect of CF on implicit knowledge to understand its role in L2 learning. There is no clear consensus on how implicit L2 knowledge should be measured and each proficiency test has its limitations (Rebuschat 2013). Therefore, it is important to measure L2 knowledge in various behaviours (Norris & Ortega 2003; R. Ellis 2005; Rebuschat 2013). For the selection of our proficiency tests, we followed the recommendations from the study by R. Ellis (2005) that compared the performance of native and L2 speakers on different proficiency tests. He found that performance on the timed grammaticality judgment task (GJT), oral production task, and an oral repetition task relied mostly on implicit knowledge. In our proposed experiment design in Chapter 2, we initially decided to use an elicited imitation task. Ultimately, for the experiments we selected an elicited oral production task instead (the Discourse Completion Task (DCT); see van de Craats (2009) for its application in a L2 study), because elicited imitation is less time efficient (only sentences that are incorrectly repeated yield information), and because it would require adaptation to high educated and low educated learners (for the experiment in Chapter 5), on account of working memory capacity, which would confound the comparison. We determined that our experiment session should not last much longer than 1.5 hours, for fatigue, logistic and recruitment purposes. The necessity of implementing two proficiency tests required that the tests be relatively fast and short.

The DCT task presented learners with a lead-in sentence, and then a sentence that they were required to complete. They were shown a picture for context, and given one or two words that could be used to make a sentence. Within 30 seconds, the learners had to record their answer (i.e. press record, speak their answer).

The GJT is one of the most often used tests in L2, but it is subject to discussion. This has to do with the level of explicit knowledge used in the GJT, but which, according to R. Ellis (2005) is dependent on the time limit: under time pressure, learners are assumed to rely more on ‘feel’ or intuition, and respond using their implicit knowledge. Another source of discussion is the nature of grammaticality. As Bard, Robertson & Sorace (1996) argued, grammaticality in a learner’s mind does not have to be dichotomous ‘incorrect/correct’. Instead, sentences should be viewed in degrees of acceptability. They argue for judgments through magnitude estimation: with learners indicating how acceptable they think the sentence is in comparison to a given stimulus, or compared to their previous
judgment. However, for participants it is difficult to rely purely on their feeling to indicate degrees of acceptability. Thus, such tasks are likely to invoke using meta-knowledge, or to reflect learner confidence and not acceptability (Sprouse 2008; Bader & Häussler 2010). To address the problem of acceptability in a GJT, we designed the task so that the participants judgments could be given on a gradient scale, but with a clear dichotomy. Learners indicated whether the sentence was correct or incorrect, but the sentence could be ‘more or less correct’ depending on where in the bar they clicked. (see Chapter 3 for a screenshot of the GJT).

The proficiency tests were analysed independently and in combination with each other. They were also correlated with questionnaire results, and with measures of learner behaviour from the logs.

5. Goals and outline of this thesis

In this thesis, we respond to the calls in the SLA literature for more controlled research on the effect of CF on oral proficiency, for more attention to individual characteristics, and for studying the learning process in addition to learning outcomes. The thesis is organized in the following way.

Chapter 2 outlines the theoretical need for systematic and controlled studies on CF to advance SLA research, and CF research in particular. We find that more research is needed on individual differences in L2 learning. This points to the necessity of developing research paradigms for CF that can take account of individual learner variation. To conduct such studies, interesting new possibilities are offered by combining ASR technology with CALL. We discuss the desirability and the possibilities of such a system that can provide individual practice, and potentially individualized practice through adaptation to learners’ needs and preferences. We present an initial experimental design which is feasible from a technological, a research, and a didactic point of view.

Following the aims for SLA research as outlined in Chapter 2, we developed and twice piloted our experiment environment that employs spoken practice with an ASR-based CALL system, GREET.

In Chapter 3, we present the system setup of GREET, the design of the speaking exercise, and how the ASR is employed to provide CF. The system is evaluated in a language learning experiment with participants with various L1s. In order to compare practice sessions and learning outcomes with and without CF, the learners were randomly assigned to one of two conditions. In one condition the learners practiced speaking with immediate, automatic CF (the CF condition), and in the other the learners practiced speaking without CF (the NOCF condition). Analysis of the proficiency gains, the logs of the practice sessions, and the learners’ evaluations of (practice with ) the system, show that GREET was successful in providing L2 speaking practice. The proficiency of the learners improved in both conditions, with spoken practice in the NOCF condition being equally effective in improving proficiency as in the CF condition. However, learner behaviour proved to be clearly different between the conditions, and learners evaluated the CF condition
more positively than the NOCF condition. As such, the CALL experiment we developed was shown to be well-suited for L2 research on oral CF. It successfully provided spoken practice with and without CF in a controlled environment, and the learners’ interactions with the system could be successfully logged and meaningfully inspected. We discuss the performance of the system, and the obtained proficiency and learner behaviour results, and relate them to current views in SLA. An important issue was the initial proficiency of the learners. An additional effect of CF was assumed to be obscured by a ceiling effect. In Chapter 4 we address this issue by recruiting lower proficient learners and closely monitoring language proficiency. 

Chapter 4 presents a detailed study on the role of CF in the development of L2 oral proficiency. This chapter presents a follow-up experiment to Chapter 3, using the same experimental setup, with improvements to GREET for usability and logging capability, and to the proficiency tests for greater reliability. We addressed the issues that were discussed in Chapter 3, and recruited learners at a lower proficiency level. This chapter also compares the two learning conditions: the CF condition and the NOCF condition. With lower proficient learners, we find differences in learning outcome, learner behaviour and learner appreciation of practice. In the analysis, we suggest ways in which learner behaviour in the CF and the NOCF condition is related to the learning outcome. Especially the differences in learner behaviour between the two conditions are discussed with references to the issues (output production, repair, and input) in the CF literature. 

A benefit of our experiment system is that it allows for comparison of multiple individual practice sessions. In both Chapter 3, and 4 we found that individuals practice quite differently with the system. 

Chapter 5 presents a study that inspects how the education level of the learners influences the effectiveness of speaking practice with and without CF. Studies on L2 learning are found to be conducted mainly with high educated learners, just as the previous chapters discussed high educated learners. There are arguments to be found in the SLA literature to assume that education background interacts with L2 learning efficiency. Because of the consistency of GREET practice, the participant sample from the language learning experiment in Chapter 4 could be expanded to include learners with different education backgrounds, from learners at primary school level upwards. The education level is seen to interact with both the CF and the NOCF conditions. This is seen for both learning outcome (proficiency gain) and learner behaviour. Similar to Chapter 4, characteristics of learner behaviour are found that relate to proficiency gain. We argue that our study substantiates that CF research with high educated learners cannot be generalized to account for all L2 learners, and that for SLA research and L2 pedagogy more effort needs to be invested in recruiting learners of various educational backgrounds. 

In Chapter 6 we discuss the findings from the preceding chapters and draw conclusions. In addition, we discuss the limitations of the current research and provide perspectives for future research.
The Role of Automatic CF in Learning L2 Grammar
Chapter 2:
Adaptive Corrective Feedback in Second Language Learning

This Chapter has been reformatted and slightly modified from:
Bart Penning de Vries, Catia Cucchiarini, Roeland van Hout & Helmer Strik (2011)

Abstract

The role of corrective feedback (CF) in second language acquisition has received much attention, and it is still a topical issue. Studies on the effectiveness of CF have produced mixed results. An essential problem seems to be that most studies on CF do not take account of individual differences, even though there are clear indications that individual characteristics influence the effectiveness of CF. This points to the necessity of developing research paradigms for CF that can take account of individual learner variation and that can adapt to the learner's needs and preferences. In this paper we suggest using a CALL system that exploits automatic speech recognition (ASR) and that is designed to adapt to individual learner differences.
1. Introduction

The term corrective feedback (CF) has been used in the literature on second language acquisition (SLA) to refer to “any indication to the learners that their use of the target language is incorrect” (Lightbown & Spada 1999). As these authors explain: “This includes various responses that the learners receive. When a language learner says, ‘He go to school everyday’, corrective feedback can be explicit, for example, ‘no, you should say goes, not go’ or implicit ‘yes he goes to school everyday’, and may or may not include metalinguistic information, for example, ‘Don’t forget to make the verb agree with the subject’” (Lightbown & Spada 1999). Further examples of different types of CF are provided in Section 2.1 below. Although intuitively one might think that correcting errors will help improve language performance, so far this has not unequivocally demonstrated in the literature. Proponents of nativist theories claim that second language (L2) learning is driven by exposure to positive evidence and comprehensible input, without any need for CF (Krashen 1985; Schwartz 1993), with Truscott (1998) taking the extreme position that CF can even be harmful. Nevertheless, there are numerous examples of L2 learners immersed in L2 surroundings, whose language proficiency stopped developing well before they fully acquired the target L2, despite ample (positive) language input (Han 2004). This suggests that exposure and input alone are not sufficient for high-level L2 learning (e.g., Swain 1985). This seems to apply particularly to adult L2 learners, who are already committed to their first language (L1) (Rohde & Plaut 1999).

Empirical studies on the effectiveness of CF have produced mixed results. This has to do in part with the fact that teachers’ CF turns out to be inconsistent (within and between studies), especially in oral (classroom) sessions of L2 learning (Carroll & Swain 1993; Iwashita 2003; Sheen 2004). For instance, teachers may correct a specific error by a learner the first time, but not react at other instances. Relevant conclusions are that CF, to be effective, should be unambiguous, consistent (Chaudron 1988), intensive (Han, 2002), and should provide opportunities for self-repair and modified output (Panova & Lyster 2002; Havranek 2002; Lyster 2004). Such conditions cannot easily be met, if at all, in natural interactions or classroom situations.

Another essential conclusion is that most studies on CF do not take account of individual differences, but discuss CF effects on groups as a whole, thus assuming that all learners use CF similarly, whereas there are indications that the effectiveness of CF is dependent on individual characteristics (see Section 3.3). The motivation for focusing on group settings and specifically on classroom settings is that these are informative about the overall educational value of CF (Lyster & Saito 2010). However, this view seems particularly tied to traditional approaches to L2 learning and ignores recent developments in CALL and language and speech technology that

---

5 Throughout this paper, L2 learning and acquisition refers to adult L2 learning and acquisition.
offer new opportunities for studying the effectiveness of CF and for implementing CF in more individualized, adaptive settings.

Can we develop a research paradigm for CF that can take account of individual characteristics and relates CF effects to those characteristics? To meet the learner’s needs and preferences, adaptive learning conditions are required to maximize CF effectiveness. In this paper we propose an adaptive approach that uses a CALL system based on speech technology. In Section 2 we provide background information on the notion of corrective feedback in second language acquisition. In Section 3 we review relevant literature on CF and indicate issues that could not be properly addressed so far. In Section 4 we explain how a CALL system with integrated speech technology can be employed to assess the role of CF in L2 learning in relation to individual learner variables. Section 5 summarizes the main conclusions.

2. CF in Second Language Learning

Nativist theories of L2 acquisition claim that language learning requires only language input, and that L2 acquisition is similar to first language (L1) acquisition (e.g. (Krashen 1985). The main arguments that CF cannot be effective according to these theories are (a) that language acquisition is an unconscious process, where CF requires conscious processing, and (b) that CF is negative evidence, whereas language acquisition is based on positive evidence, i.e. language input. Theories in favor of CF call either of these arguments into question. Though L1 acquisition can be assumed to be largely an unconscious process as infants lack the cognitive development required (Hulstijn 2005), it is not conclusively shown that CF (which is consciously processed) is irrelevant for L1 acquisition (e.g., Saxton 1997). Moreover, even if conscious learning does not benefit L1 acquisition, the further developed cognitive abilities of adults may still enable consciously processed CF to benefit L2 acquisition. Even more importantly, empirical evidence abundantly shows that adult L2 learners often do not reach a high level of L2 proficiency. That means that there are essential differences in acquiring a L1 and a L2. Even if CF is not effective at all in L1 acquisition, it may be effective in L2 acquisition.

2.1. Types of corrective feedback

Lyster and Ranta (1997) distinguish six types of feedback in their often-cited classroom observation study:

- **Explicit feedback**: teacher provides the correct form and clearly indicates that what the student said was incorrect.
- **Recasts**: teacher reformulates all or part of a student’s utterance, minus the error.
The Role of Autom
catic CF in Learning L2 Grammar

- **Clarification request**: teacher formulates a question indicating that the utterance has been unclear or ill-formed and that a repetition or reformulation is required.
- **Metalinguistic feedback**: the teacher response contains either comments, information, or questions related to the well-formedness of the student’s utterance, without explicitly providing the correct form.
- **Elicitation**: teachers try to elicit the correct form by asking for completion of a sentence, or asking questions, or asking for a reformulation.
- **Repetition**: teacher repeats the erroneous utterance in isolation.

Types (2) and (6) provide implicit feedback: it is up to the learner to notice that an error was made. The other types are explicit in indicating that an error occurred. The categories are not as clear cut as may seem, since there may be degrees of explicitness or implicitness. Additionally, CF may be accompanied by visual cues or intonation. Moreover, the interpretation of the distinction relates to the setting of the feedback, e.g., implicit recasts may be argued to be explicit in formal classroom settings (Lochtman 2002). The exact definition of what CF type is studied is often lacking in studies (see Nicholas, Lightbown & Spada (2001) for a discussion), which complicates comparing and synthesizing results Norris & Ortega (2000).

### 2.2. Explicit and implicit knowledge and learning

Two types of language knowledge are distinguished in the literature: explicit knowledge and implicit knowledge. Explicit knowledge refers to knowing rules, while implicit knowledge involves using a rule without being aware of the rule. Explicit knowledge is open to conscious inspection, whereas implicit knowledge is not. The distinction is related to the one between declarative knowledge and procedural knowledge in cognitive psychology (Anderson & Fincham 1994). The acquisition of cognitive skills has a declarative and a procedural stage. Processing in the declarative stage is slow and controlled, whereas processing in the procedural stage is fast and beyond active control. Through frequent use, the facts and rules acquired in the declarative stage become automatic procedures.

As explained by DeKeyser (2007) “automatized knowledge is not exactly the same as implicit knowledge. While implicit knowledge or implicit memory is always defined with reference to lack of consciousness or awareness … absence of awareness is not a requirement for automaticity”. He further explains how it is possible to have knowledge that is implicit but not automatic, for instance when learners make many errors and fluency is low, and to have knowledge that is automatic but not implicit, in the case of fluent learners that make few errors but are still conscious of rules, for instance language teachers or linguists. Although declarative and procedural knowledge are not exactly the same as explicit and implicit knowledge, they are highly similar and in this paper they will be treated as such for practical purposes.

Closely related, but also clearly different, is the distinction between explicit and implicit learning. Where explicit learning involves conscious intention to find
out whether the input information contains regularities or specific elements, implicit learning is input processing on an unconscious level without such an intention (Hulstijn 2005). Krashen (1985) argues that language proficiency is based on implicit knowledge of the language, and that only implicit learning can increase the learner’s proficiency. Implicit learning takes place, he writes, when a learner is exposed to comprehensible language input. Since CF appeals to explicit learning, it is assumed to have no effect on L2 development.

Though there is agreement among SLA researchers that language proficiency is based on implicit knowledge, there is disagreement on whether and how explicit knowledge can contribute to language proficiency. The core issue is whether explicit knowledge stays separate from implicit knowledge, or whether there is an interface and some kind of exchange.

2.3. Interface

Krashen (1985) argues that there is no interface between implicit and explicit knowledge. Proponents of the weak interface position (N.Ellis 2005) argue that conscious learning can, under specific conditions, further the acquisition of implicit knowledge. If the learner’s attention is explicitly directed to features of the L2, this knowledge will first be stored as explicit knowledge, but it may under specific circumstances become implicit, for instance by (repetitively) applying explicit knowledge in production (De Bot 1996). Proponents of the strong interface position argue that implicit knowledge can gradually be formed from explicit knowledge, without any further constraints (DeKeyser 2003).

2.4. Noticing

Schmidt’s Noticing Hypothesis (Schmidt 1990) claims that, for learning to take place, a learner must be consciously aware of the difference between his/her rendering of the L2 and the target L2. However, some features of the target language may be difficult to perceive in native speech (e.g. phonetically reduced, or semantically redundant morphology) (N.Ellis 2002), causing learners not to hear and notice them in their input. In interaction, for example, the conversation partner may only break the flow of conversation to correct an error if the meaning of the speaker is not understood. While the level of conscious attention for noticing is debated (e.g. N. Ellis 2005), the assumption that noticing a feature in the input is a first step in language acquisition is shared by several researchers (see Cross 2010 for an overview). This suggests a potentially important role for CF to facilitate noticing and focusing learner attention on errors and correct L2 forms. Since exposure to L2 will not automatically guarantee this kind of noticing, CF must come into play to draw the learner’s attention to language-specific and individual problems and (indirectly) stimulate them to attempt self-improvement (Long 1996; Havranek 2002).
2.5. Learner differences

The outcome of L2 learning may vary among learners and can result both in high proficiency and low proficiency speakers. Other than L1 acquisition that leads to uniform success, it seems that contextual factors and individual differences have a greater influence in L2 learning (R. Ellis 2010) than in L1 learning.

In an overview of individual differences in L2 acquisition, (Dörnyei 2005) identifies five main concepts that predict L2 learning success: motivation, personality, aptitude, learning style, and language learning strategies. Additional variables such as anxiety, creativity, learner beliefs, self-esteem, and willingness to communicate are mentioned but are said to fall under the aspects of personality (anxiety and creativity under motivation or personality), or are in need of more research to determine their exact role and nature. On another level, age, gender are also found to influence L2 learning (Lyster & Saito 2010). Though clearly age and gender are learner differences, they are separate because they interact with each of the five psychological factors. Learner differences are good predictors of L2 success according to (Dörnyei 2005), but at the same time the characteristics are difficult to pin down in an exact definition. This complicates the theoretical discussion of these variables, but in (educational) practice these factors have been found to influence learning. The importance of individual differences is currently widely recognized in educational contexts and “a great deal of research has been conducted in educational psychology on how to adapt instruction to the strengths weaknesses and preferences of the learners” (Dörnyei 2005 ). In light of the influence of learner differences on L2 learning, CF, and the type and manner in which CF is provided, is likely to affect each learner differently (see for a discussion of CF types, Section 3.3.). Therefore, it seems that L2 learning could be improved if the learner received optimal CF; in other words, CF adapted to the learner’s characteristics (Cronbach & Snow 1977).

What should be noted, however, is that since many learner characteristics have an influence on language learning, it also becomes necessary to take into account the context of language learning. The setting and situation of language learning will interact with learner characteristics (Dörnyei 2005; Lochtman 2002) and should be considered.

3. Research on corrective feedback

As explained above, theories of L2 learning make different predictions about the role of CF in language acquisition. The theories have had their influence on language pedagogy (e.g., Krashen’s influential comprehensible input theory caused many language teachers to focus mainly on communication), but in turn, pedagogical practices inform theories on the effect of instruction on L2 learning (e.g. Long 1983). Despite many empirical findings with respect to CF effectiveness, these findings are difficult to combine to support one particular view on the role of
In this section we look at CF research so far, and the reasons why CF research has not yet provided conclusive results.

### 3.1. Overview of CF research

In a meta-analysis of 49 studies on language instruction (classroom, laboratory, interaction studies), Norris and Ortega (2000) found that these studies collectively suggest that instruction has a positive effect on L2 learning; CF was included as a type of instruction. However, in their discussion, they raise several issues with respect to research methodology, one important issue being that several experiments were found to test for explicit knowledge, yet made claims about implicit knowledge. They make several recommendations for improving research practice, such as to reduce the number of variables and to design experiments with replication in mind.

Several meta-analyses of CF studies (Lyster & Saito 2010; Russell & Spada 2006; Mackey & Goo 2007) suggest that CF is effective, and that explicit CF is more effective than implicit CF (types of CF are discussed below). However, caution should be taken when interpreting these results due to different definitions, different operationalizations, and different measures of CF effectiveness adopted in CF studies (Lyster & Saito 2010; Nicholas et al. 2001). Additionally, determining which type of knowledge (i.e. explicit or implicit) is responsible for a learner’s performance on pre- and post-tests remains problematic (R. Ellis 2005), as well as establishing to what extent learning effects are durable beyond the post-test period (Doughty 2003).

Another problem for CF research is variation between individual learners and teachers. For instance, in classroom settings, teachers may have difficulty delivering CF following specific linguistic targets in a consistent manner (Nicholas et al. 2001). Arguably, therefore, studies of CF produce varying results, because “in real classrooms, students rarely get much, if any, individualized attention, and corrective feedback, if provided, is usually given ad hoc, covering a wide range of interlanguage constructions” (Han 2002).

Overall, then, research suggests that CF is potentially effective for L2 learning. However, the studies fail to be conclusive due to variability between studies, the variables under study, and uncertainty of outcome measures. Studies on CF should therefore aim to deliver CF in a controlled manner, and be rigorously defined so that replication of the study is possible, to enable cross-comparison and synthesis across experiments.

### 3.2. Effects of different types of corrective feedback

The exact definition of what type of CF is studied is often lacking in studies (see Nicholas et al. (2001) for discussion), which complicates comparing and synthesizing results (Norris & Ortega 2000). In observation studies of CF in the classroom, recasts turned out to be by far the most frequent technique for error
The Role of Automatic CF in Learning L2 Grammar

Correction (Havranek 2002; Lyster & Ranta 1997; Lochtman 2002). Compared to the more explicit ways of giving CF that halt the conversation to point to language form (e.g., L: ‘the man goed to the market’, T: ‘No, that is incorrect. The man went’), the recast causes less learner anxiety and does not disrupt the flow of communication (L: ‘the man goed to the market’, T: ‘The man went to the market’). However, precisely the quality of being so discrete causes recasts to be often unnoticed as CF (Nicholas et al. 2001). As a result, though most frequently used in the classroom, a recast may not be the most effective type of CF.

Clarification requests, metalinguistic clues and elicitation (types 3, 4, 5 from Section 2.1), are so-called negotiation of form techniques, collectively called prompts. They indicate that an error was made without providing the correct form. These feedback moves are considered to be effective because they induce learners to reprocess their output, and to produce “pushed output” (Swain 1985; De Bot 1996), but have been criticized because they would contribute to explicit linguistic knowledge and not to competence.

An important factor that turns out to mediate the effectiveness of the various CF types is research setting. In general, laboratory studies indicate greater effectiveness of CF than classroom studies, probably because important variables such as intensity and consistency are better controlled for in laboratory studies (Lyster & Saito 2010). Furthermore, there are important learner characteristics that may be connected to the relative effectiveness of different feedback moves and that so far have received little attention in CF studies (Sheen 2010; Lyster & Saito 2010), as will be discussed in the next section.

3.3. CF and individual differences

Very few studies in CF research address learners individually. Classroom observation studies such as Lyster and Ranta’s (1997) are very informative about the type of CF that is most used in pedagogical settings (see above), but do not provide information about the specific input and output of an individual learner, and how the individual differences interact with CF (DeKeyser 1993). Experimental studies also often use groups that receive a particular CF treatment (e.g. Mackey & Philp 1998). However, individual and contextual factors are likely to play a larger role in L2 acquisition than in L1 acquisition. It seems necessary, therefore, to pay more attention to individual differences (R. Ellis 2010). In an individual setting, it is easier to adapt CF to the learner. An example that individualized CF can be more effective is found in Han (2001). Here a learner was seen to consistently misinterpret written CF from a tutor. Only after specific CF that was adapted to this learner’s developmental readiness and L1 background did the learner effectively correct her grammar. Additionally, there are indications that while some learners may require very explicit and immediate forms of CF, others do not appreciate being interrupted during conversation, thus preferring more implicit and delayed feedback moves (see e.g. Strik, Cornillie, Colpaert, van Doremalen & Cucchiarini 2009).
This general neglect of individual differences and preferences may be related to an equally general feeling that L2 learning research should first of all provide guidelines for educational approaches, which up to now, have been mainly classroom oriented. Recent technological developments provide new opportunities for more individualized, tailor-made approaches to L2 learning in which adaptation can play a prominent role, as will be explained in the next section.

4. New opportunities for implementing and investigating CF in L2 learning

Although it is difficult to draw firm conclusions on the role of CF in SLA, several requirements stand out in the findings discussed above that are necessary to establish effective CF. For each learner, CF must be unambiguous, understandable, detectable, and short (El Tatawi 2002, Russell 2009). These criteria can only be met when individual learner differences are taken into account. What is needed is individualized attention, a consistent focus over a longer period on one type of error only, intensive treatment, and, finally, consideration of the learners’ developmental readiness and learning style.

These demands are hard to realize in classroom settings where teachers have to distribute their attention over a group of learners and where there is no direct control over what a learner is paying attention to. CALL systems with integrated speech technology may create settings where individual learner variation can be taken into account. How can we use the new technological opportunities to create the settings where learners receive CF adapted to their needs and preferences, and are provided with sufficient opportunities for self-repair and modified output (El Tatawi 2002)?

4.1. CALL systems

More optimal learning conditions can be created by resorting to a CALL environment where CF is provided individually. To date, different CALL systems have been used for experimental research on second language acquisition (see, e.g. Hulstijn 2000), in particular to study the role of CF (e.g. Bull 2000; Sachs & Suh 2007). CALL systems offer several advantages for research on CF. Learners interact one-on-one with the system which provides immediate, clearly defined, consistent feedback on all learner utterances. The CALL system engages learners more intensively, and can motivate them to practice until they achieve a ‘perfect score’ (Wachowicz & Scott 1999). Additionally, scores on tasks and developmental progress can be logged by the computer. This gives the possibility of optimizing the CF and adapting the program to individual characteristics, e.g., developmental readiness. For research purposes, all data, such as learner output, and reaction times in response to tasks can be logged and analysed.

The CALL systems used so far use written input, even when investigating oral skills (Payne & Whitney 2002; Sagarra 2007). However, writing is assumed to
employ explicit L2 knowledge (Bialystok 1981) while to study the role of CF it is necessary to target implicit knowledge through on-line performance as in speaking. This could be achieved by employing CALL systems that make use of Automatic Speech Recognition (ASR), a specific application of speech technology, intended to parse the incoming speech signal into words. This parsing task is not trivial, especially in the case of non-native speech (van Doremalen, Cucchiarini, Strik 2010). Developing high-quality ASR-based CALL systems that can properly handle L2 speech requires the combination of different types of expertise. This might explain why this approach has not been adopted earlier.

4.2. An ASR-based CALL system for studying the effect of CF on oral proficiency

In the project FASOP (Feedback on Acquisition of Syntax in Oral Proficiency) we aim to develop a learning setting that copes with individual variation, is replicable, and targets on-line processing by recognizing and analysing spoken output. To this end we will use a CALL system that employs ASR to analyse the spoken output of learners of Dutch as L2 and to provide systematic, clear, consistent, and adaptive CF on syntax in oral Dutch L2 performance. In this section we motivate and describe the setup of our system for studying CF.

4.2.1. Advantages of ASR implementation in a CALL system

Developing the proper adaptive ASR-based CALL system requires an interdisciplinary approach. We need to combine and integrate knowledge from the fields of pedagogy, teaching Dutch as a second language, research methodology, and speech technology, in order to increase our understanding of the role of CF in L2 learning and of the effectiveness of adaptive CALL systems.

In this system learners engage individually in dialogues with a virtual language tutor, receive CF on incorrect utterances, and are stimulated to produce modified output. Learners will provide spoken answers. By using an ASR system that is trained on non-native speech, and by constraining learner output, L2 utterances can be recognized with high precision. The FASOP project employs the technology developed in the project ‘Development and Integration of Speech technology into Courseware for language learning’ (DISCO) (Strik et al. 2009).

CF is individualized in the sense that it builds on the errors made by a specific learner. The CF targets particular linguistic items or addresses various related linguistic features at the same time. Moreover, learners can practice as long as they want, at their own pace. This provides additional opportunities for studying the effect of CF on L2 learning, by varying focus, intensity and length of treatment. The crucial advantage for CF research is that all these factors can be handled with a degree of control and systematicity that is not achievable in traditional classroom situations. The effect of different feedback types can be studied by providing CF
through various feedback moves and comparing their effectiveness (see Section 4.2.2.).

A further innovative feature of the FASOP project is that the CALL system will adapt CF on the basis of learner responses and successes. In particular, it adapts to the learner by providing the type of feedback that appears to be more effective. Only a very few studies have investigated the impact of adapting feedback moves in relation to learner characteristics and achievements.

An obvious advantage of applying an ASR-based CALL system is that all learner-system interactions can be logged, to enable a thorough analysis of language input, corresponding learner’s spoken output, the system’s feedback and the learner’s response to the feedback. In the FASOP project these different data sources provide together the basis for the systematic investigation of the role of CF in L2 learning. These sources will be complemented with data derived from pre- and post-tests in which more traditional proficiency tests are administered, supplemented by expert judgments on learners’ performance. In this way, a comprehensive picture of the impact of CF and its long-term, generalizing effects can be obtained.

A final advantage of this approach to CF research is that it can provide guidelines for improving L2 teaching. Just as traditional, group oriented research was considered to be informative for traditional classroom-oriented L2 education, individualized, adaptive research on L2 learning can provide useful insights and guidelines to develop L2 teaching methods that are individualized and adaptive. In other words, this type of research also has educational value and the experimental conditions are ecologically valid because if the system developed in laboratory conditions appears to be effective, it can also be applied in real learning situations, as a supplement to traditional lessons.

4.2.2. Measuring the effect of CF

One of the problems in CF research is measuring effectiveness. As mentioned in Section 3.1 in (Norris & Ortega 2000) observed that several of the studies they examined used a pre- and post-tests that could be said to target explicit language knowledge. Since the FASOP project specifically focuses on progress in language proficiency, it is necessary to target implicit knowledge. During treatment, this will be done mostly by having learners produce spoken language output (and possibly implementing some form of time pressure, which is another effective way of minimizing the influence of explicit knowledge on production (R. Ellis 2005).

Crucially, the pre- and post-tests must test the level of implicit knowledge, to determine whether the treatment in which CF was provided had an effect. For this reason, we use a timed grammaticality judgment task (GJT) and an elicited imitation (EI) task (see Vinther (2002) and Erlam (2006) on EI tests; see Tremblay (2005) for an overview of GJT). In the EI test, a learner attempts to verbally repeat specific sentences of the L2. The goal of the test is that the sentences exceed the capacity of the learners’ working memory, so that they cannot repeat the sentences verbatim. In restructuring the sentence, the test is able to show where problem areas of the
The Role of Automatic CF in Learning L2 Grammar

learner’s grammar lie. The GJT is a test in which learners have to decide whether a sentence is grammatical or not in the L2. To target implicit knowledge it is essential that this test includes time pressure, so that the learner does not have time to reflect on the sentence but must respond intuitively. These tests were found to be most indicative of implicit language knowledge in a study by R. Ellis (2005), in which he compared test results of native and non-native speakers on various types of proficiency tests.

In FASOP, we will look at the performance and acquisition of learners on one particular grammatical structure, namely verb placement in Dutch. This increases the level of control that we have on the experiment (effectiveness of CF type is also seen to interact with grammatical structure in R. Ellis 2007). Additionally, the effect of the treatment can easily be examined in the pre- and post-tests as performance on the target structure.

Information on learner characteristics will be collected through a questionnaire. The data from the questionnaire can then be compared with effectiveness of types of CF. Both performance data during the treatment of the experiment (rate of speech, reaction times, self-correction after CF) and performance of the learner on the pre- and post-tests will be examined for correlations.

4.2.3. Design of planned ASR based CALL experiment

The FASOP experiments will use an ASR-based CALL system that engages the learners in a dialogue with a virtual tutor. After watching a short film clip, the learner receives questions about the video, to which the learner must respond by recording a spoken response. Due to the high level of precision we require of the ASR and because the ASR must analyse non-native speech, it is necessary to constrain the dialogue (van Doremalen et al. 2010). The responses of the learner must be predictable. As a result, we have decided to provide the learner with ‘blocks’ with which they must construct the sentence, e.g.:

Q: ‘Waar gaat dit filmpje over?’ (what’s this film about?)
Block 1: ‘dit filmpje’ (this film)
Block 2: ‘gaat’ (is)
Block 3: ‘over twee jonge mannen’ (about two young men)
Block 4: ‘die verhuizen’ (moving house)

This form of constraining the spoken output by the learner enhances the possibility that the produced utterance is correctly identified by the speech recognizer. The feedback that the system gives will be either, a) implicit recast, b) explicit recast, or c) prompt. These are operationalized as follows:

a) Implicit recast: neutral background colour plus the correct form of the learner’s utterance
b) Explicit recast: red background colour, the statement ‘no, that is incorrect’ and the correct form of the learner’s utterance
c) Prompt: red background colour, the statement ‘no that is incorrect. Try again’. The learner then can re-record the utterance. If the learner again makes an error, the system fills in one block and asks for a learner repetition, until the learner produces a correct utterance.

In this fashion, we can examine two characteristics of CF: implicit (implicit recasts) versus explicit CF (explicit recasts, prompts), and CF that provides the correct form (implicit and explicit recasts) versus CF that stimulate self-correction by the learner (prompts). However, recasts can hardly be said to be implicit in a computer system that is specifically designed to practice L2. Learners using the system are likely to be focused on language form, and not on meaning. To counter this problem, we also include some multiple choice questions about content of the film. For instance, if the man in the film buys apples at the market, the system may ask ‘what did the man buy?’ and the learner must respond by constructing a sentence, e.g., ‘the man bought two apples/pears at the market’. A CF move by the system may then be interpreted by the learner as feedback on content, and not on form. Additionally, the system will also randomly repeat correct sentences to further ensure that the implicit feedback is implicit to the learner (see also Section 2.1).

In this manner it is possible to systematically provide CF on spoken learner utterances using ASR. By comparing the performance of a control group that works with the system but does not receive CF from the system, with the experimental groups, we can determine in a highly controlled experimental environment whether CF is effective for learning a specific grammatical structure of Dutch. Additionally, by examining the effectiveness of CF types, and relating that to learner characteristics, it is possible to adapt the CF according to learner requirements.

5. Conclusions

We have argued that a new research paradigm is needed to study CF in second language acquisition in which clear, systematic, consistent, intensive and adaptive CF can be delivered to language learners. ASR-based CALL systems can provide such a paradigm. Further experiments are required to support our approach and to provide evidence for the role of CF in second language acquisition.
The Role of Automatic CF in Learning L2 Grammar
Chapter 3:  
Spoken Grammar Practice and Feedback in an ASR-based CALL System

This Chapter has been reformatted and slightly modified from:

Abstract

Speaking practice is important for learners of a second language. Computer assisted language learning (CALL) systems can provide attractive opportunities for speaking practice when combined with Automatic Speech Recognition (ASR) technology. In this paper we present a CALL system that offers spoken practice of word order, an important aspect of Dutch grammar. The system uses ASR technology to process the learner’s responses and to detect errors so that immediate corrective feedback (CF) can be provided on learner errors. We evaluate the system as a learning environment by analysing proficiency gains in pre- and post-tests, the logs of the practice sessions, and the learners’ appreciation of the system. In this paper we present two learning conditions: 1) in one condition the learners received oral practice and immediate CF on spoken performance, and 2) in the other condition learners received oral practice and NOCF on spoken performance. We found that our system was successful in providing L2 speaking practice. Results show that both groups improve their proficiency on the target feature as a result of treatment. Between the groups there is no significant difference in learning, but the groups proceeded differently through the sessions, and the learners in the group that received automatic CF evaluated the system more positively than the NOCF group. We discuss the performance of the system as an environment for language learning, and the obtained proficiency test results, and relate them to current views on second language acquisition.
1. Introduction

Improving second language (L2) spoken proficiency requires practice speaking the language. Speaking provides learners with opportunities to notice gaps in their L2 knowledge (Swain, 1985), to increase automatization of forms by using them in production (De Bot, 1996), and to improve skill-specific performance (DeKeyser, 2007). However, opportunities to practice speaking in an L2 classroom are limited. Besides the limited time available per learner, it is impossible for a teacher to tailor speaking exercises to the needs of each individual learner (Truscott, 1999). Providing learners with more opportunities to practice speaking in the L2 is therefore an important goal for L2 pedagogy.

Opportunities for speaking also exist outside the language classroom of course: immersed learners have possibilities for interaction with native speakers, and foreign language learners can access the internet for practice facilities. Analyses show that these types of interactions are focused primarily on meaning and not on accuracy of the output (Lyster, Saito & Sato, 2013), and that feedback does not always come from a trusted authority (Dickinson, Eom, Kang, Lee, & Sachs, 2008). Since spoken proficiency is determined by fluency, accuracy and complexity (Housen & Kuiken, 2009), it is important to combine the practice of producing spoken output with focus on linguistic form (cf. comprehensible output, Swain, 1985; Ellis & Bogart, 2007).

Against this background it is interesting to look at CALL systems, as they offer the promise of unlimited individual practice, and can provide automatic corrective feedback (CF) on learner output by implementing automatic speech recognition (ASR). Potentially, ASR-based CALL systems can supply excellent conditions for speech practice, due to reduced anxiety for L2 speaking (Warschauer, 1996), the possibility to practice forms intensively, and the control learners have over their learning environment, e.g. timing, duration and pace of their practice session (Torlakovic & Deugo, 2004). Implementing ASR in a CALL system, however, presents specific challenges for technology and the pedagogical design. Perhaps as a result of this, there are currently no systems that offer focused practice and automatic feedback on grammar in the spoken modality (Bodnar, Cucchiarini & Strik, 2011). To address this gap, we developed an ASR-based CALL system for spoken grammar practice. In this study we evaluate the performance of our system by using it in two different learning conditions, spoken grammar practice with and without immediate automatic CF.

2. Research background

This section provides the background for the development of our CALL system, and discusses why a CALL system for ASR-based practice of spoken grammar is feasible and desirable. In this overview, we pay particular attention to the role of spoken output and the role of CF in adult L2 learning.
2.1. Spoken grammar practice through computer assisted language learning

Several studies have shown that CALL can be effective for language learning (see for a review Golonka, Bowles, Frank, Richardson, & Freynik 2012), and may even outperform classroom instruction (e.g., Torlakovic & Deugo 2004). The reasons for its effectiveness are availability, the amount of individual practice, and the learner’s level of control over the learning experience. Practice with a computer is also found to result in reduced anxiety about making mistakes, which can be a restraining factor in the classroom or in face-to-face interactions. The anonymity provided by computer chat results in learners reporting that they feel less stressed, and that they produce more, and more varied, output (Warschauer 1996).

On the whole, CALL systems mostly address written production, but the beneficial characteristics of CALL mentioned above can also apply for speaking practice in CALL (cf. Eskenazi 2009). The modality of practice is relevant, as spoken production requires a higher cognitive load and control over the articulatory system (Payne & Whitney 2002). To train control over cognitive load and articulation, CALL systems should allow for speaking practice and provide automatic corrective feedback on speaking performance. Finally, there is a call for studying the effectiveness of CALL in the “relatively unexplored” field of “speaking online” (Felix 2008, p. 157).

The advances in the field of ASR technology make it possible to implement focused exercises for spoken production (Clifford & Granoien 2008). In a review of 350 studies that report on the effectiveness of CALL systems, Golonka et al. (2012) found strong evidence that ASR technology can impact pronunciation training, and that those ASR systems provided feedback effectively. They do not report systems that provide automatic CF on spoken grammar. A review study of CALL systems using ASR by Bodnar et al. (2011) showed that the available systems address communicative skills or pronunciation, while there are currently no systems that offer spoken practice with automatic CF on grammar errors. Since accurate grammar is an important aspect of proficiency, and therefore a main pedagogical goal in L2 learning, a system to practice grammar in the oral modality would be a valuable application for L2 learning.

2.2. Language production and second language learning

In our study we take the interactionist view on second language acquisition (SLA), which assumes that interaction and the accompanying feedback serve an important function in developing the interlanguage (Pica 1994; Long 1996; Gass 1997). The roles of output and of CF have been subject to debate in the SLA literature. As they are relevant to the current study, we briefly review them here.
2.2.1. The role of output

The role of output in L2 learning has been amply discussed in the SLA literature, with most of the evidence pointing toward a positive effect of output (Murano 2007). A theoretical basis for the relevance of language production in SLA is outlined in the Output Hypothesis (Swain, 1985). The central idea is that during output processing learners have to construct their message to convey the intended meaning, and must do so using grammatical structure. In this way, producing output gives rise to several ways that can draw the learners’ attention to gaps in their interlanguage.

The Output hypothesis contrasts with the Input hypothesis (Krashen 1982), which states that L2 acquisition is driven by comprehensible input. Though it is evident that language input is necessary for SLA, Swain (1985) shows that it is not always sufficient. This is the main drive for the Output Hypothesis. An explanation of why input alone could be insufficient is given by Izumi (2003), who discusses the relation of comprehension and the need for accurate syntactic processing. For comprehension, learners can rely on additional information that is not necessarily linguistic (e.g. contextual, pragmatic, logical). This enables them to comprehend the meaning of a message, even when their interlanguage grammar cannot accurately parse the message. As a result, deficiencies in the learners' interlanguage do not necessarily surface during input processing.

The Output Hypothesis indicates three possible functions of output: 1) a noticing function, 2) a hypothesis testing function, and 3) a metalinguistic function. In referring to skill-acquisition theory (Anderson & Fincham 1994), De Bot (1996) added the function of enhancing fluency through practice, thus stressing the importance of producing spoken output to improve speaking proficiency.

2.2.2. The role of corrective feedback (CF)

Though the role of CF in SLA is still controversial, a considerable body of research has shown that CF has a positive effect on language learning (e.g., Norris & Ortega 2000; Russell & Spada 2006; Lyster & Saito 2010). CF can be important because it can draw attention to (the presence of) an error in a learner's interlanguage. In this way it can cause the noticing of features in the L2 that have not been acquired yet, or have been acquired inaccurately. The relevance of attention is incorporated in Schmidt’s Noticing hypothesis (1990, 1994), which states that conscious noticing of form features in the L2 is facilitative for them to be acquired. Noticing can occur without CF, when the learner attends to language forms in the input (cf. VanPatten 1994) or when producing output (Swain 1985). However, it is possible that certain features (e.g. grammatical form-meaning relationships) are not noticed by the learner (N.Ellis 2006), and that CF is required to draw the learner’s attention to those features. Focus on language form through CF is therefore likely to be beneficial for language learning.
Despite many studies on CF, it is not clear under which conditions CF is effective (Lyster et al. 2013). Factors that are emerging as influencing CF effectiveness are educational setting, type of CF (Lyster & Saito 2010), and learner differences (Dörnyei 2005). Controlled research is required to further investigate these issues (Russell & Spada 2006; Goo & Mackey 2013).

In an overview of CF research, El Tatawi (2002) lists the conditions for CF to be effective for interlanguage development: CF needs to be systematic and consistent, clear enough to be perceived as CF, it should allow for time and opportunity for self-repair and modified output, the intention of CF should be clear, and the learner should be ready for the feedback. This overlaps with the objections Truscott (1999) raises against grammar correction in oral practice; he claims that it is impossible to achieve the ideal characteristics of individualized CF in classroom environments and therefore argues for abandoning the practice of grammar correction altogether. However, Truscott (1999) suggests at the same time that “the possibility remains that some untested combinations of these variables could produce successful feedback, while avoiding (or minimizing) the accompanying problems” (p. 451). Through the individual approach in a CALL system, it may be feasible to start exploring the possibilities for successful CF.

2.3. Research questions

We find that it is important to provide L2 learners with ample opportunities for skill-specific practice, in order for them to improve their spoken proficiency. This is best done through form-focused output practice (pushed output), since fluency and accuracy are important features of proficiency. A way to provide learners with more opportunities to practice under said conditions is to make use of an ASR-based CALL system. We designed a system that provides learners with the opportunity to practice grammar in the spoken modality. The system interacts with the learners and provides CF on grammatical accuracy by employing ASR technology. This places demands on our design since it requires that the ASR accurately processes non-native speech from different L1 backgrounds to provide pedagogically beneficial CF. Although our design takes into account the limitations of ASR, an important point in our evaluation is to test to what extent the system is effective in improving learners' proficiency. To this end, we set up a language learning experiment in which two learning conditions were implemented, one based on practicing with spoken output and opportunities for self-correction (NOCF group), and the other based on practicing with spoken output and immediate automatic CF (CF group). This way we can evaluate the working of the system and compare its performance in these two learning conditions. Thus we address the following research questions:

- Is it possible to develop an ASR-based CALL system that can detect grammatical errors in spoken performance and provide appropriate corrective feedback?
• Is there a difference of effect between practicing with spoken output and self-monitoring (NOCF group), and practicing with spoken output and immediate automatic CF (CF group)?

3. Method

This section describes our ASR-based CALL system that offers grammar practice in spoken production. The learning exercises we developed are embedded in an experiment to measure learning outcomes by applying pre- and post-tests, and learner appreciation through a post questionnaire. During the exercise learner behaviour is logged.

3.1. The GREET system architecture

In Figure 1, we give a schematic overview of the practice system. The learner interacts with the system through the graphical user interface (a screenshot is given in 3.5, Figure 2), and is presented with a task from the courseware database. The learner is given a question about a film clip, accompanied by ‘word blocks’ to form a spoken response. The ‘word blocks’ restrict the number of possible responses and this contributes to higher ASR accuracy.

For each question in the exercise, the language model contains all answer possibilities (i.e. all possible sentences) that can be created with the word blocks. These are tagged with meta-information, which in the current experiment is simply whether that answer-sentence is grammatically correct or incorrect. In later experiments, this setup can include more detailed messages.

When the learner records an utterance, the speech recognizer outputs a recognition result and a confidence level. If the confidence level is below a pre-set threshold, the system concludes that the learner did not record a valid attempt and the learner is asked to re-record. In other cases, the recognition result is mapped to an answer sentence in the language model, and the system sends to the courseware engine the appropriate message regarding the sentence’s grammaticality (for more detail see Strik, Cornillie, Colpaert, van Doremalen & Cucchiarini 2009).

The final step is the presentation of a feedback message to the learner. Taking the output of the error detection step as input, the courseware engine determines what type of feedback to present, and how it should be presented (see section 3.5.1).

The experiment is run through a website and the learner interacts with the system through a web browser. All recognition is performed on a central web server. This allows us to run more experiments at the same time and at different locations, making use of existing computer lab facilities, and to store all data centrally.
Figure 1. A schematic overview of the GREET system.

In the present experiment the system behaves differently in the two conditions. For both groups, the speech is processed through the recognizer, and the recognition result (and confidence) is logged. For the CF group the grammaticality of the ASR output is evaluated and immediate feedback is shown on the screen. The learners advance or retry depending on the system’s response. The learners in the NOCF group receive no immediate feedback on their performance and they advance or retry according to their own preference.

A benefit of a CALL system is the logging of system-learner interactions (cf. Collentine 2000) to allow for insight into learner behaviour. Throughout the experiment, our system stores learner-system interactions, allowing us to look in detail at learner behaviour. Relevant for the current paper are the interactions with respect to each question, the number of attempts per question, and whether CF led to a correction in case of a learner error.

3.2. Participants

We recruited 29 adult participants from Dutch language courses at A2 and B1 (CEFR level) at the Radboud In’to Languages centre in the Netherlands. They were offered 15 Euros for participating. The participants were randomly assigned to the CF group, or to the NOCF group. Random assignment and drop-outs resulted in unevenly populated groups: twelve participants in the control group (6 male, 6 female), and seventeen in the experiment group (5 male, 12 female).

In total, there were 14 different L1s in our sample: Arabic, Chinese, Dari, English, French, German, Indonesian, Italian, Russian, Luganda, Polish, Portuguese,
Romanian, and Spanish (NOCF group: 7 L1s; CF group: 12 L1s). The participants had received over eight years of formal education after primary school, with most at university level. Their mean age was 31 years with a range of 22 to 48 years old.

3.3. Procedure

Participants signed up for two sessions of one and a half hours, at times of their choosing. Each session consisted of 45 minutes of various tasks and 45 minutes of spoken grammar practice. In the first session, they completed a questionnaire on personal data, two pre-tests (described in section 3.5), and then moved on to do a treatment session (section 3.4). Within a week of the first test day they completed the second session with the second part of the treatment, two post-tests, and finished with a post-treatment questionnaire (section 3.6).

The participants logged in and the website gave them their task in a step-by-step fashion. Each time they completed a task, a new one was shown. Before the proficiency tests, the participants were given instructions presented as a slideshow, and then completed three trial questions before moving on to the real test.

3.4. Target structure

The syntactic feature under investigation is inversion of subject and verb. This feature affects word order in Dutch. Dutch is a verb second (V2) language, and V2 entails that the finite verb appears in the second position regardless of the first constituent in the main clause. The subject precedes the finite verb when it is the first constituent (see (1) below). The order is inverted when another constituent precedes the finite verb (see (2) below). The subject then immediately follows the finite verb.

The acquisition of V2 is problematic for L2 learners of Dutch (Jordens 1988). Violation of V2 may remain in the interlanguage of L2 speakers for a long time. This is because it is not necessarily 'noticed', as it does not affect the meaning of the sentence, and may therefore not be salient in input and output. This makes V2 an appropriate feature to study the effect of CF, as CF is likely to benefit errors that do not affect meaning, do not typically lead to communication breakdown, and that lack a clear form-meaning relationship (N.Ellis 2006; Sauro 2009).

6 In our sample, we have one participant whose L1 (German) has V2 (all Germanic languages have V2, except for English (Haider 1985)). Bohnacker (2005) found that V2 transferred for Swedish learners (a V2 language). Interestingly, for Swedish learners who had English as their L2 (non-V2), there was less transfer: suggesting influence of their L2 on learning a new language. In our sample, the German participant had English as an L2, but showed no problems with Dutch V2. Excluding this informant from the analysis did not change the outcomes in any relevant way.
3.5. Treatment design

Participants practiced implicitly with the target structure. They watched short clips (approximately 35 seconds) of a story about a man moving to Amsterdam and interacting with his neighbours, developed by an educational publisher (Put & Peekel 2001). After each clip, an on-screen teacher asked 3 to 5 questions about the content. One third of these questions were designed to elicit the target structure; none of the questions themselves contained the target feature. To answer, the participants had to construct a sentence using ‘word blocks’ (see Figure 2): parts of a sentence that need to be combined to form one sentence. The participants recorded their answer by speaking into a microphone.

The treatment was divided over the two experiment sessions, for 45 minutes of practice in each session. In total there were 120 questions (63 in session 1; 57 in session 2), of which 40 were target questions (21 in session 1; 19 in session 2). We controlled for time-on-task, so that if the participant completed all the questions of the particular session before the 45 minutes were up, they looped back to start at the first sentence of that session.
The Role of Automatic CF in Learning L2 Grammar

Figure 2. A screenshot of the treatment exercise. The teacher, Greet, asks a question, and the learner answers by pressing the record button and speaking. Answers are formed with the blocks in the second row (starting with ‘naar de …’). The first block of an answer is sometimes given to require a particular grammatical structure, e.g. for target questions. As a hint after negative CF, a word block will be moved onto a ‘?’-block to show its correct position. The example translates as: ‘Melvin is going to move. What does he do first?’ The blocks translate as: Eerst ‘First’, gaat ‘goes’, Melvin ‘Melvin’, naar de gemeente ‘to city hall’. (The teacher character is originally from the DISCO project (see: Strik, van Doremalen, Colpaert & Cucchiarini 2013), used with permission).

3.5.1. CF in treatment

We want to measure the effect of CF by comparing two groups that both produce output, but where only one receives CF on their output. However, CF can be given in various forms: it can differ in degrees of implicitness and explicitness, and can be classified as either input providing (reformulations) or output pushing (prompts) (see for an overview Lyster, Saito & Sato 2013, p. 4-5). These characteristics are all assumed to mediate the effectiveness of the CF types (see R. Ellis (2010) for a discussion). For the purpose of group comparison, we required a CF type that was most likely to be effective for language learning.

In a meta-analysis of studies comparing CF types, Lyster and Saito (2010) conclude that prompt feedback is the most effective. Prompt CF is an explicit
indication that the utterance was incorrect, but withholds the correct form so that the learner must attempt to self-correct. Its effectiveness is attributed to the characteristics that prompts are explicit in stating that there is an error, which facilitates noticing (Ammar 2008), but since they do not provide the correct form, they induce self-processing in the learner (Panova & Lyster 2002) and trigger self-repair, thereby pushing the learner to produce modified output. In line with skill-learning theory, uptake following prompts and containing repair enables learners to strengthen their control over linguistic forms that have been only partially acquired (Lyster et al. 2013). Though the CF that GREET provides is different from oral CF in the classroom (cf. Heift 2004), we designed the CF in our system to share these salient characteristics of prompt CF.

In GREET we operationalised CF as follows:

- After an incorrect answer, a red pop-up is shown containing the message “That is incorrect. Try again”. The learner has to retry, but receives a hint from the system. To form an answer, the learner uses the randomly ordered word blocks (see Figure 2): the system can give a hint by placing one of the word blocks in its correct place. In this way, the system elicits the correct answer from the student by incrementally adding sections of the answer sentence (i.e. the word blocks).
- If the ASR could not confidently process the learner’s response, a light-grey pop-up was shown, containing the message “I’m sorry, I did not understand you. Please try again”, which we coded as DNU feedback. No hints were given.
- After a correct (i.e. grammatical) answer a green checkmark pop-up was shown, and the learner was automatically advanced to the next question.
- For the NOCF group the message was the same regardless of their answer: a white screen saying “Your answer has been saved. You can only save one answer. Do you want to keep this one and move to the next question, or try again?”, after which the learner had the choice to progress, or retry.

The learners in the NOCF group did not receive immediate CF from the system, but they were told that their speech would be analysed by the system and that they would receive a score at the end. This was done because in our pilot experiments the subjects in the NOCF group were reluctant to complete the experiment. We had noticed that while the subjects in the CF group would complete both sessions, subjects in the NOCF group often did not return for the second session, and indicated that they felt they were not practicing effectively because they did not receive feedback from the system.

3.6. Proficiency tests

Two proficiency tests were selected to measure knowledge of the target grammatical feature (accuracy). We selected two tests for cross-task comparison and validation (Norris & Ortega, 2003): a timed grammaticality judgment task (GJT) and a discourse completion task (DCT). The tests were selected based on the psychometric study by R. Ellis (2005). The tests measure the same aspect of language competence,
but are distinct in that one is a receptive reading task, and the other a spoken production task.

3.6.1. **Grammaticality judgment task (GJT)**

In the GJT participants judged 40 sentences, with a time limit for the judgment of each question set at 12 s based on pilot versions. Each test had an equal number of target and filler sentences, and grammatical and ungrammatical sentences were also equally distributed (see appendix A for items).

Pre- and post-test versions were counterbalanced, and the order of item presentation was randomized per subject. Correct judgments scored 1, incorrect judgments scored 0. A response outside the time limit was scored as incorrect.

---

**Figure 3. An example of a GJT item.** The learner is instructed to read the sentence and indicate on the bar whether the sentence is correct or incorrect in Dutch. The item above can be translated as: In the store Laura sees her friend.

3.6.2. **Discourse completion task (DCT)**

The discourse completion task (DCT) elicits oral production. The target structure under investigation is easily avoided in Dutch, so we restricted the possibilities for output, and modified the design to make postverbal subjects obligatory in target sentences. Participants saw the beginning of a sentence which they were required to complete. To establish some context for the task, they were given a lead-in sentence, one or two hint words, and a picture (see Sanz and Morgan-Short (2004) for a
similar task design for written production). To answer the participant pressed the record button and spoke a full sentence. In this task, there was a time limit of thirty seconds. The test was counterbalanced, and the order of item presentation was randomized for each participant. Each test version was made up of 32 items of which half were targets, and half were fillers (items in appendix B). The recordings were transcribed and scored for correct use of V2.

Figure 4. An example of a DCT item. The learner is instructed to complete the unfinished sentence, using the context provided. Answers must be spoken and recorded. The item above can be translated as: Wat doet Kim om twaalf uur? Om twaalf uur... boterham + eten. ‘What does Kim do at twelve o’clock? At twelve o’clock... sandwich + eat’.

3.7. Post-test questionnaire

After the test, the participants were asked about their opinion of the GREET system. Subjects indicated on a five-point Likert scale whether they agreed or disagreed with statements about the experiment system. Questions asked if they felt that their level of Dutch had improved, their self-confidence to speak Dutch had grown, whether it was a good system for learning Dutch, and whether they had enjoyed using it (see appendix D for the items). They were also asked open questions to elicit opinions and suggestions about the system and CF.
4. Results

4.1. System performance

In a pilot study we determined the CF accuracy rate in our system, which was found to be high (96% correct accepts, and 97% correct rejects (Bodnar, Penning de Vries, Cucchiarini, Strik & van Hout 2011). In the current experiment, we examined several indicators such as practice logs and questionnaire responses to verify that the CF provided by the ASR was as expected.

During the sessions, the 29 participants practiced a total of 1176 target questions: the CF group (N=17) practiced 708 questions (M: 41.65 SD: 15.99), and the NOCF group (N=12) practiced 468 questions (M: 39.00, SD 10.72). Since the system only interacts with the CF group, we examined their practice session logs. A good indicator of system performance is the number of attempts per question. Ideally, a learner is guided to the correct answer with help of CF within a maximum of 4 attempts, when the full answer is given by the system. However, when the system cannot confidently process the learner’s response, it outputs ‘did not understand’ (DNU) feedback. Because in this case the system does not give a hint, the number of attempts can exceed 4. This happened 29 times out of 708 questions practiced by the CF group. We can allow some cases where this happens, because on inspection these would frequently include learner recording errors (microphone use, truncating of recorded audio, not pronouncing all the given blocks), and then the DNU feedback is often justified.

There were 6 instances where a question elicited too many attempts (more than 8), and the participant was unable to proceed to the next question. This would cause a disruption because assistance of the experimenter was necessary. These 6 instances were divided over 5 female participants with different L1s. Inspection of these cases revealed that a combination of factors caused these problems. For 4 of these cases, there was a problem with the sound files that were sent to the recognizer. This resulted in distorted sound, and the recognizer could not confidently process the signal. A combination of the sentence content, the speaker, and the sound error, prevented the recognizer to provide CF on these 4 particular sentences. The sound error seems to have been related to the wireless connection. In another instance, the participant had placed the microphone incorrectly, which resulted in bad recordings. In the final instance, the participant did not consistently use all the blocks, which, in combination with pronunciation problems, confused her on how to answer.

Overall, the system performed well in the CF group, by providing CF adequately on 679 out of 708 questions (95.9%), which we take to be an acceptable number. We found 6 severe disruptions and 23 possible disruptions in the experiment flow, which may affect the experiment and disruptions like these will be addressed in a future experiment.

After excluding the 6 problem instances, the average number of attempts per question was 1.56 (SD 1.093). The sentences with the highest average (sentence
number 41 (means 2.41, SD 1.80) and sentence number 126 (means 2.94, SD 1.89)) turned out to be challenging: 41 was a marked construction, 126 contained a series of difficult vowels (tense vs. lax distinctions, diphthongs). At the same time these two sentences were challenging for the ASR, as the word blocks contained phonologically similar content, which resulted in a lower recognition confidence, and thus a higher number of DNU feedback.

In total, 12 of the 17 participants in the CF group completed the full experiment without assistance from the experimenter. They only received help with the login and startup. This means they could progress through the practice session with the system independently. For 5 participants in the CF group, there was a disruption in their treatment session. These were the 6 cases mentioned above where the learner could not proceed. We experienced one server crash which caused a disruption. In the NOCF group, 12 participants proceeded at their own pace, and did not experience disruptions.

4.2. Proficiency tests

As a first step in our analysis, we inspected the item scores in both proficiency tests. The internal reliability scores (Cronbach's alpha) of the GJT was 0.86, and for the DCT 0.95. The two tests showed a high correlation (pre- plus post-test data: r=.781, p=.000). However, the correlation is not perfect and the tests measure competence in a different modality, so we take the tests to provide us with complementary information on the learner’s language level. Figure 5 shows the data from the pre- and post-tests.
The Role of Automatic CF in Learning L2 Grammar

Figure 5. Bar graph showing the mean score (proportion correct) of the two proficiency tests and their standard errors. The CF group received immediate CF after each utterance; the NOCF group had the possibility to re-record their answer after each utterance.

In a t-test we found no significant difference between the groups at the pre-test, (GJT: t (27) = -2.03, p = .841; DCT: t (27) = -.569, p = .574) nor at the post-test (GJT t (27) = .211, p = .834; DCT t (27) = .760, p = .454). To test the improvement in both groups we applied an ANOVA, repeated measures with group (CF vs. NOCF) and treatment (pre vs. post) as independent variables. All effects had an F value < 1, with the exception of a main effect of treatment. For the GJT, we found a significant effect (F (1, 27) = 5.209, p = .031), and for the DCT a nearly significant effect (F (1, 27) = 3.856, p = .060). To test if non-target items were learned, the GJT-fillers were analysed. No effects were found: including no effect for treatment (F (1, 27) = 1.501, p = .231), and no interaction of treatment with group (F (1, 27) = .341, p = .564). This indicates that the learning occurred only for the target items, and that subjects have improved on their accuracy of Dutch V2 word order as a result of treatment. It also means that no evidence of an interaction of group with treatment was found. This
indicates that the CF group did not have an additional learning gain as a result of prompt feedback.

Some participants scored high on the pre-tests, suggesting a ceiling effect. As these subjects cannot learn significantly in either condition, they may obscure the CF and NOCF comparison. Therefore we analysed a subset of our sample, only including participants whose pre-test score could improve by more than two items. We re-ran the analysis on this subset (N=25; CF (N= 14), NOCF (N=11)), but did not obtain results substantially different from those reported above.

4.3. Log data

The system logs interactions with the participant, which can be inspected for information on learner behaviour during the experiment. In the two sessions combined, there are 40 unique target questions (21 in the first session, and 19 in the second). After completing all of the questions of a session, the learner restarts at the beginning, so that all learners practice for the same amount of time. This results in much individual variation in the sessions with respect to the number of questions practiced. In 90 minutes of training (two sessions) we find that there is a range of 9 - 69 target questions practiced (range CF-group = 9 - 69; NOCF group = 26 - 66). Some participants, then, do not practice all target questions: they do not complete either of the sessions (9 in the CF group, 5 in the NOCF group), or they complete only one of the sessions (2 in the CF group, and 3 in the NOCF group). At the same time, there are participants who complete both sessions (6 in the CF group, and 4 in the NOCF group), with four participants repeating a significant number of questions. Though all participants practice the same amount of time, the amount of input received, and output produced during practice varies. As a result we looked more closely at learner behaviour and proceeded to analyse the difference in learning behaviour between the CF and the NOCF group.

The logs show us clear treatment differences. In the bar graph in Figure 6, the data for the two sessions are shown. Figure 6 shows the number of questions practiced (Q), and the number of attempts (A), where a learner can have several attempts per question. When a learner makes an error, there is chance to repair the error in the subsequent attempt. The number of times this happens successfully is the number of repairs (R).

---

7 One participant in the CF group practiced only 9 target questions. This is much lower than the next lowest with 18 questions practiced. This participant had serious pronunciation difficulties, aggravated by the distorted sound problem. As a result she had many instances of DNU feedback, resulting in two disruptions in her session. We included her in the analysis as she did practice for 90 minutes (and produced many attempts). In the post-test, we found that her proficiency scores had not improved.
64 | The Role of Automatic CF in Learning L2 Grammar

Figure 6. Bar graph of GREET treatment sessions, with standard errors. It shows the number of questions practiced (Q), attempts (A), and repair (R), per session.

The number of questions practiced is the same for both groups, in session 1 (Q1: t(27) = -.407, p = .687) and in session 2 (Q2: t(27) = -.467, p = .644). The CF group makes significantly more attempts. When comparing the NOCF and the CF group, we found that there is a significant difference in the number of attempts in session 1 (A1: t(27) = -4.285, p = .000), and in session 2 (A2: t(27) = -3.027, p = .006). The CF group also had significantly more counts of successful repair in both sessions: session 1: (R1: t(27) = -5.496, p = .000), session 2 (R2: t(27) = -4.666, p = .000).

Since we find much variation in the number of questions practiced, we examine how learner behaviour correlates with proficiency scores, and pre- post-test gain in those scores. As participants in the two conditions interact differently with the system, we split this analysis per group. The NOCF group does not show significant correlations. In the CF group, pre- and post-test scores are positively correlated with the number of questions practiced in 90 minutes of treatment (pre-test: r = .554, p = .021; post-test: r = .601, p = .011); and negatively correlated with the number of repairs (pre-test: r = -.810, p = .000); post-test: r = -.778, p = .000), and number of attempts per question (pre-test: r = -.580, p = .015; post-test: r = -.590, p = .013). This suggests that the CF condition provides more assistance in the form of CF for less proficient learners. This is not the case in the NOCF condition, where the
log data of the learners’ sessions is not correlated with learner proficiency. Interestingly, we do not find any significant correlations of log data with pre- post-test proficiency gains in any of the two conditions: neither when including or excluding the ceiling participants. The amount of output or input during practice is not found to correlate with proficiency gains.

4.4. Post-test questionnaire

In a post-test questionnaire we asked the participants about their practice session in 17 questions, of which 12 were identical for the CF and NOCF group. The questions were designed to gauge learner appreciation by asking them about enjoyment, usability and effectiveness of the GREET system. Participants responded using a 5-point Likert scale (1= very negative, 5=very positive), from which we calculated their evaluation of the system. The data is represented in Figure 7 below.

![Figure 7](image)

**Figure 7.** Bar graph showing participants’ evaluation of the GREET system in the post-test questionnaire, with standard errors. Responses were given on a 5-point Likert, with 1= strongly disagree, and 5 = strongly agree. The asterisk indicates that the questions were negatively phrased (i.e. a score of 5 would be negative): here the score has been inverted for easier visual interpretation.
The mean for both groups is above 3, which shows that the participants were moderately positive about using the system (3 = neutral). In a reliability analysis of the questionnaire, we find that Cronbach’s alpha (N=12) = .765. This suggests that the questions elicited sufficiently reliable responses. On the individual questions, we see in Figure 7 that the CF group is consistently more positive than the NOCF group. In order to determine whether there is an overall group difference, we averaged questionnaire responses, and we found that the CF group is significantly more positive about the system, as they show a higher appreciation score (t(29) = 2.372, p=.025).

5. Discussion

To address the first research question, we find that the GREET system that we developed managed to offer spoken L2 practice of a specific feature of Dutch grammar. It successfully processed the learners’ spoken output, identified errors and provided corrective feedback. The output practice in the constrained design of our grammar training appeared to be effective for learning in both conditions, CF and NOCF. The ASR system was able to deal with a large range of L1 backgrounds. Additionally, the individual variation in the logs of questions practiced showed that learners progress at different speeds through the exercise, and that the system could provide learning at a pace that is appropriate to the proficiency levels of the learner.

Another important point is that the learners were positive about practicing with the system. Moreover, the group that practiced with CF, and received the most interventions by the system, was more positive than the NOCF group.

Overall, 23 out of 29 learners could proceed without problems through the program, making it a good experimental setting (excluding 5 participants with ASR problems and 1 participant where the server crashed), as there was minimal influence from the experimenters, and it was consistent across participants.

However, there were some issues that caused interference in our language learning program. Issues with speech recording errors could be countered by familiarizing the learners with the recording interface before the experiment, and giving them a chance to listen to their own recording and reporting any problems. The sound error problem can be addressed by using only stable internet connections, and better recording technology.

The second research question was whether there is a difference in learning gains between practicing with spoken output and self-monitoring (NOCF group), and practicing with spoken output and automatic CF (CF group). We found that both the CF group and the NOCF group showed improved accuracy on the tests. Though we found a larger number of utterances (attempts) and repairs in the logs of the CF group, this did not lead to a higher overall improvement of proficiency, which we expected based on previous research on the effect of CF on the development of L2 grammar (Sauro 2009; Sagarra & Abbuhl 2013). Apparently, the NOCF group managed to learn from practice as much as the CF group without receiving immediate corrective feedback. An explanation may be the higher cognitive
involvement in task completion by the NOCF group. Since these learners had to monitor their own answers, they had to pay more attention to the input and to their utterances, especially because they were told that their speech would be evaluated and scored by the system. Though the input learners receive in our treatment does not contain the target feature, there may have been an effect of task essential practice, as assumed by input processing (VanPatten 1994) (cf. Sanz and Morgan-Short 2004 who find practice with structured input equally effective as practice with explicit rule information). In any case, learners may have strongly monitored their own output, a favourable learning condition according to the output hypothesis. This explanation is corroborated by the observation that the number of questions completed in the CF and NOCF group was the same. The NOCF group answered the questions at the same pace as the CF group, suggesting expended effort in trying to answer correctly. The limitation could be that output practice without CF is effective only when the learner has prior knowledge of the grammatical structure practiced in the treatment, since it relies on the learner noticing their error (cf. Ammar & Spada 2006, p. 563). A closer investigation of the pre-test proficiency scores seems to indicate that the NOCF group participants with a low entry level were unable to improve their accuracy on the target structure. This suggests that while learners with higher proficiency levels may learn without CF, those with lower proficiency may require CF to improve their grammatical accuracy. This is in accordance with other reports in the SLA literature that show a relationship between proficiency level and the effect of CF (Ammar & Spada 2006; Russell & Spada 2006). Given the limited number of subjects in this experiment, these findings should be handled with caution because they concern only two learners. However, they suggest a line of investigation to be pursued in future experiments.

A related issue for CF effect is the target feature that was practiced. Nagata (1996) argues that the effectiveness of CF in their CALL system may have been related to the complexity of the structure, where more complex features benefit more from CF. However, the target feature in our study, V2 inversion, is not an easy task to learn. It is a persistent error in L2 Dutch, occurring even in advanced learners. From the learners’ point of view, it is a difficult feature to master (cf. Blom & de Korte 2011). Nevertheless it would be interesting to investigate learning with CF of other grammatical structures in our system to shed light on this issue.

Another relevant factor that should be taken into consideration is that the NOCF group was told that their answers were recorded and scored by the system. This might indicate that a sense of interaction and monitoring by the system is sufficient to stimulate learners and push them to make the most of the input (cf. Chapelle 1998). Against this background, a valuable contribution of ASR in our system seems to be creating that sense of interaction that learners apparently need by letting them feel that something (i.e. the CALL system) is listening to their utterances and judging their quality.

Important information on the role of CF in this study was also obtained from the post-questionnaires. These revealed that the learners valued receiving CF from the system, regardless of the impact this had on their performance. This may tie in with the idea that learners appreciate the feeling of interaction: the immediate
The role of automatic CF in learning L2 grammar

response of the system on their utterance. Relevant may also be the role of the positive feedback in the form of a green check mark. It is not uncommon that learners indicate a preference to receive CF. Lyster et al. (2013) summarize research that indicates that, in general, learners want to receive CF to a greater extent than teachers perceive (p. 8). In this respect, ASR-based CALL systems are well suited for providing consistent and intensive CF. Moreover, it is easy to imagine that ASR-based CALL systems can give learners control over the amount of CF, and type of CF that they receive (cf. Heft 2004).

Other possible explanations for the lack of a significant effect of CF on grammar development in this study are related to the relatively short treatment and to the specific type of CF employed. Indeed, the subjects practiced and received CF for a relatively short time, only ninety minutes, and it is possible that providing CF for a longer period might be more beneficial (Lyster & Izquierdo 2009, p. 485). As to the type of feedback provided in this study, elicitation, a prompt indicating that the utterance was wrong and eliciting a reformulation from the learner, we explained that we chose this feedback move because previous studies on the effectiveness of CF have indicated the advantages of explicit feedback. However, in future experiments we intend to investigate other feedback moves, also in relation to other factors such as learners’ L2 proficiency levels, educational levels and preferences.

6. Conclusions

In this paper we have shown that it is possible to develop an ASR-based CALL system that can detect grammatical errors in spoken performance and provide appropriate CF. We succeeded in creating a successful learning environment in which we can monitor the learner’s behaviour during practice.

We can conclude that our system is effective since the learners improved their grammatical accuracy as a result of practice with the system. We obtained significant learning gains for a group of 29 participants practicing for ninety minutes with our system. In the group that received CF, the ASR component of our system successfully interacted with learners with twelve different L1s.

Using the GREET system we researched whether there is a difference of effect between practicing with spoken output and self-monitoring (NOCF group), and practicing with spoken output and immediate automatic CF (CF group). The answer is negative for learning gain. We did not find a significant additional effect of immediate, automatic CF on grammar development. The ASR component still seems to serve an important function however. First, it enables us to provide learners with a sense of interaction that may be necessary in the conditions that we investigated where spoken output is required. Second, the group receiving ASR-based CF evaluated the system more positively than the group that did not receive CF. The learners in the CF group were confident that the system was a good way to learn Dutch, and that their Dutch had improved as a result of working with the system. This suggests that our current setup is a good way of practicing grammar for
oral proficiency and that even if CF may not be strictly necessary for improving oral grammar, it may still be requested and appreciated by L2 learners.

In this evaluation of the GREET system we have shown that ASR-based CALL systems can be a valuable addition to language learning pedagogy, by providing an opportunity to practice L2 speaking in a controlled and individual environment. The system we have developed is particularly suitable for conducting experiments on the effect of different CF types under various learning conditions and with different groups of learners. It also means that such a system will support investigating the role and impact of individual learner differences.

7. Acknowledgements

We would like to thank our colleague Joost van Doremalen for developing the ASR component of the CALL system used in this experiment. This work is part of the research program ‘Feedback and the acquisition of syntax in oral proficiency’ (FASOP), which is funded by the Netherlands Organization for Scientific Research (NWO). We are indebted to three anonymous reviewers for providing helpful suggestions to improve the quality of the manuscript.
Chapter 4:
Effect of Corrective Feedback for Learning Verb Second

This Chapter has been reformatted and slightly modified from:

Abstract

This paper presents a detailed study on the role of corrective feedback (CF) in the development of second language (L2) oral proficiency. Learners practiced speaking with a computer-assisted language learning (CALL) system that employs automatic speech recognition (ASR) technology to provide CF. The system tracks learner behaviour by logging the system-user interactions. Two language learning conditions are compared. In the CF condition learners received immediate, automatic CF on the grammaticality of their spoken output. In the NOCF condition, learners practiced speaking with the option to self-correct. The target structure under investigation is Dutch verb second (V2) in the main clause. The results show that learner proficiency improved in both conditions. The CF condition shows an additional benefit for learning that is related to the learner’s initial knowledge of the target structure (which we call V2 proficiency). Learners at a lower V2 proficiency level benefitted more from practice with CF than learners in the NOCF condition. Learner evaluations are in line with these results: both the CF and the NOCF groups positively evaluated practice with the system, but the CF condition was preferred by learners starting at a lower V2 proficiency level. For more information on these outcome measures, we investigated the learners’ behaviour during practice. The two groups were found to receive equal amounts of input, but learners in the CF condition produced more (grammatically correct) output during treatment. We found that the CF group repaired their errors in fewer attempts as they progressed through practice. Learners in the NOCF condition generally did not (attempt to) repair their errors. However, the learners answered correctly more often as they progressed in the training. The log data, therefore, shows learning of the target structure in both conditions. We discuss these results and how learning outcome is related to learner behaviour.
1. Introduction

Research so far indicates that corrective feedback (CF) contributes to adult second language (L2) learning. Multiple meta-analyses of studies have found an overall positive effect of CF on L2 learning (e.g. Russell & Spada 2006; Lyster & Saito 2010; Li 2010). However, between the studies in these meta-analyses, CF effectiveness varies. This is because many variables appear to influence the effect of CF (Russell & Spada 2006; Ellis 2010). Examples of such variables are (a) learner-external variables: learning context (Sheen 2004), type of error that CF targets (Ellis 2007), type of CF (Lyster & Ranta 1997; Lyster 2004), and (b) learner-internal variables: individual differences such as learner proficiency (Mackey & Philp 1998), working memory (Mackey et al. 2002), motivation (Uzum 2010), anxiety (Sheen 2008). This has led several researchers to call for more detailed and controlled research on factors influencing CF (Russell & Spada 2006; Ellis 2010; Ellis & Sheen 2006; Lyster et al. 2013; Goo & Mackey 2013). The focus in CF research, therefore, is shifting from whether CF is effective, to what determines its effectiveness (Mackey & Gass 2006; Ellis 2009).

To find out what determines CF effectiveness requires knowledge on how individual learner differences interact with contextual CF variables. This can be obtained only through experiments that are tightly controlled, with a rigorous methodology, with neatly operationalized variables, necessary for replication studies. In a computer-assisted language learning (CALL) setup, many variables can be controlled while learners interact individually with an exercise. Most notably the input and the system’s reactions to the learner are controlled: the behaviour of the ‘tutor’ is consistent. This makes CALL well-suited for CF research. Although much CF research concerns spoken production (Sheen 2010), CF research using CALL has mainly been limited to the written modality. However, by incorporating Automatic Speech Recognition (ASR) in a CALL environment, it is possible to study the effect of CF on oral production (Penning de Vries, Cucchiarini, Strik & van Hout 2011). With an ASR-based CALL system it is possible to run experiments in which learners practice speaking the L2, individually and with consistent CF, with controlled L2 input, and where all learner-system interactions are logged for analysis of the learning process. This way, these L2 learning experiments are replicable.

In a previous study we presented the ASR-based CALL system that we developed (Penning de Vries, Cucchiarini, Bodnar, Strik, & van Hout 2014), and showed that it was successful in providing automatic CF on spoken grammar and that it improved learner proficiency. However, providing CF on spoken practice did not yield an additional learning gain over spoken practice. We assumed that the participants’ high level of proficiency regarding our target structure reduced the effectiveness of CF, or that it increased the effectiveness of spoken practice. For the few learners who scored low on pre-test proficiency, there were indications that learners with CF improved, whereas learners without CF did not. The added effect of CF, therefore, seemed to be related to the learner’s initial proficiency level.
SLA and CALL research underscore the importance of looking at both the product of acquisition, and the process of acquisition, to obtain a more comprehensive view of the language learner and the learning process (Mackey & Gass 2006; Schulze 2011). Relevant processes that take place during L2 learning, such as noticing and interaction with input and output (cf. Schmidt 1995; VanPatten 2004; Swain 1985; DeKeyser 2007) can be investigated by examining learner behaviour (Hegelheimer & Chapelle 2000; Chapelle 2005, 2007). In addition, this is necessary for a pedagogical design: to design maximally effective language learning tasks, it is important to know what learners are doing during the task (Fischer 2007). SLA research shows that individual differences mediate L2 learning and the effectiveness of CF. If tasks and CF are sufficiently adapted to the learner, learning results can be improved (e.g. Han 2001; Doughty & Long 2003). This adaptation is desirable for pedagogy, and CALL systems make it feasible (Chapelle 2007; Heift & Schulze 2007). However, further research is required to determine how logged data is related to learning outcomes. Developing CALL systems that automatically determine when and how to adapt to the learner, requires knowledge of how processes during practice relate to efficient learning. Thus systems that link learner behaviour with proficiency gain can inform language pedagogy, by offering grounded results on when and how to correct errors. In addition, they add to a more detailed understanding of when and how CF plays a role in L2 learning, and allow for empirical testing of assumptions made in the SLA literature on learner-internal processes during language learning.

In this paper we present a study on CF in which we look in detail at the learning outcome, input, and process, and how they may be related. We first situate our research in the SLA and CALL literature and formulate our research questions in Section 2. In Section 3 we describe our experiment system (GREET) and method. In Section 4 we present results of the proficiency tests and log data analyses. We discuss our research findings and end with a general discussion in Section 5.

2. Research background

This section reviews the relevant literature on how CF affects language learning (2.1), and studies that investigate how learner characteristics, specifically learner proficiency, mediate CF effectiveness (2.2), and how learner behaviour in CALL exercises can be recorded and related to language learning outcomes and processes (2.3). On the basis of this review we present our research questions in Section 2.4.

2.1. Language learning and CF

L2 learning through input and meaning based interaction results in varying levels of L2 attainment (Swain 1985). To improve L2 learning efficiency, it is necessary to draw the learner's attention to language form, for instance through instruction or CF (Doughty & Long 2003). Relevant in this respect is the 'Noticing Hypothesis' by
Schmidt (1990, 1995) that posits the necessity of (sub)consciously noticing language features as a prerequisite for learning.

Noticing of form can happen at different instances during interaction. Learners can notice language form in the input on account of its saliency or frequency (VanPatten 2004; N. Ellis 2002). Learners can also notice forms in their L2 output (Swain’s 1985, 1995 Output Hypothesis). During production, learners may notice ‘gaps’ in their language knowledge when they are unable to express themselves sufficiently in the L2. In addition, the production of language forms aids the process of automatization (De Bot 1996; DeKeyser 1998): forms that are stored as declarative knowledge may be accessed more rapidly over time when they are practiced, and production of language forms contributes to the development of procedural knowledge, a result of which more fluent speech may evolve. Practicing language forms in the appropriate modality is important according to theories of skill acquisition (DeKeyser 2007): i.e. to improve speaking proficiency, learners need speaking practice. According to these theories, producing spoken output can improve L2 proficiency.

CF can help trigger the processes of noticing and proceduralization: it can point out errors in the output, provide (modified) input (e.g. as a result of negotiating for meaning), and prompt reformulations of errors. Various taxonomies of CF types have been proposed, but overall they all specify that (a) CF can be explicit or implicit in pointing out the error, on a gradient scale, and that (b) CF can provide input, and/or be output pushing (Lyster et al. 2013). As a result, the operationalization of CF provided is relevant (cf. Nicholas et al. 2001).

As a result of receiving CF, the learner may produce modified output. Lyster and Ranta (1997) pointed to the possible relevance of modified output as indications of CF effect. So far, however, the role of modified output has been disputed. If the modified output contains repair of the error, this may indicate that the learner noticed the error, but this inference is problematic (Ammar & Spada 2006; Hegelheimer & Chapelle 2000). For instance, repair following a recast may be the result of mimicking (Gass 2003). Production of additional output can be beneficial from a skill-acquisition point of view. However, a contrasting view (by e.g. Long (1996) and Gass (1997)) argues that the CF facilitates the learner’s noticing of forms in the input through interaction but that the uptake is not relevant to CF effectiveness. Loewen (2004), in an analysis of 32 hours of meaning-focused lessons, found that uptake after interactional CF predicted learning of features on subsequent post-tests. On the other hand, Mackey and Philp (1998) show that CF leading to most modified output is not necessarily the most effective CF. The relationship between uptake and learning has been a source of discussion (Ammar & Spada 2006; e.g. see the discussion between Lyster & Ranta (2013) and Goo & Mackey (2013)), and so far there seem to be very few studies investigating the link between uptake and learning (Loewen 2004: 161-162).
2.2. CF and learner proficiency

Individual differences mediate language learning and the effectiveness of instruction (Ehrman, Leaver & Oxford 2003). Moreover, individual learners vary in the ways they react to, and benefit from interactional CF (e.g. Mackey 1999; Mackey et al. 2002; Mackey et al. 2003; Sheen 2004, 2008; Swain and Lapkin 1998). This suggests that CF may be individualized to improve learning results (cf. Cronbach & Snow 1977; R. Ellis 1994), an example of which is seen in Han (2001). Since many variables interact to determine CF effectiveness, it is necessary control learner external variables as much as possible in order to research how the learner internal variables mediate CF effect. Here we focus on proficiency level of the learner.

Several studies suggest the importance of the learner's proficiency level for CF effectiveness. In (Penning de Vries et al. 2014) we found indications that the level of proficiency as measured on the pre-test regarding a specific target structure influenced the effectiveness of CF. In a study of eight adult L2 learners Lin and Hedgcock (1996) found that the effect of CF types is different at various stages of interlanguage development. Ammar and Spada (2006) compared the effectiveness of two CF types, prompts and recasts, in a sample of 64 young L2 learners (grade 6). For high proficient learners, the types were equally effective, but low proficient learners benefited more from prompts (explicit CF). In a similar comparison with 55 L2 learners (mean age 20 years) recruited at university, Iwashita (2003) found that only high proficient learners could benefit from positive evidence in interaction, but both low and high proficient benefited from explicit CF. Li (2013), in a study with 78 university L2 learners (mean age 21 years), found that explicit metalinguistic CF was more effective than recasts (implicit CF) for low proficient learners, while the two feedback types were equally effective for high proficient learners. To summarize, low proficient learners are found to benefit more from explicit CF, whereas CF types seem equally effective for high proficient learners (though see Ellis, Loewen and Erlam 2006 for a discussion on the issues in research on the effectiveness of CF types). Another factor that is closely related to proficiency is whether the CF provided is appropriate to their proficiency level. Learners are found to respond best to developmentally appropriate recasts (Mackey & Philp 1998; Philp 2003). However, providing learners with developmentally appropriate CF requires continuous adapting to their instructional needs as their proficiency improves, and this requires knowledge of when the next aspect of the language should be taught and in which order (R. Ellis 1994).

Learner proficiency also plays an important role in how much a learner notices in the L2 (Mackey et al. 2002). Hanaoka (2007) finds that more proficient learners notice more of their errors in output. It is necessary, therefore, that “the forms that are noticed lie within the learner’s ‘processing capacity’” (Ellis 2000: 8). For language practice with, and without CF, proficiency level influences L2 learning.

For studies on the impact of proficiency level on L2 learning, the level of the learner needs to be specifically defined. For example, within proficiency levels as defined by the CEF-level, learners may differ on their mastery of particular aspects
of the L2. It is therefore necessary to look at individual proficiency on specific target structures (Shintani et al. 2014: 108). To illustrate, for spoken interaction, Li (2013) found that the effect of CF interacted with the target structure and the learners’ proficiency level. Li provided recasts and metalinguistic feedback on L2 learning of Chinese classifiers and perfective. Recasts were equally effective for low and high proficiency level learners for classifiers, but for the perfective, recasts were only effective for high level learners. Metalinguistic feedback was found to be more effective overall for low level learners, but both CF types were equally effective for high level learners. Another possibility is to make sure that learners were never exposed to the L2 before (Yilmaz 2012), or to employ an artificial language (De Graaff 1997). However, this severely limits the possibilities of studying CF effects on acquiring new features of the target language.

Though the proficiency level of the learner is assumed to influence CF effectiveness, further research is needed as to how this effect varies in relation to other learner differences. For instance, learners at (approximately) the same proficiency level interacted with CF differently according to age and interlocutor in Mackey et al. (2003). Moreover, most CF research is conducted with high educated learners (Young-Scholten 2013), and research is needed to determine whether these findings generalise to low educated L2 learners. For this reason it is necessary to inspect in detail the individual results and their interactions with CF.

2.3. Learner behaviour

SLA perspectives such as sociocultural, interactionist, and dynamic systems theory emphasise the necessity of taking the learning process into consideration besides learning outcomes (Aljaafreh & Lantolf 1994; Mackey & Gass 2006; De Bot et al. 2007). In addition, SLA researchers have criticized approaches that rely on single measures of proficiency, for instance using single pre-and post-tests (Hulstijn 1997; Norris & Ortega 2003; Mackey 2006), and recommend measuring learner behaviour in a more general, encompassing way for cross-validating of learning results.

A valuable contribution of CALL to SLA and pedagogical research lies in its capabilities of tracking learner behaviour (Fischer 2007; Collentine 2000; Hulstijn 2000; Chapelle 2007). This has been applied in several CALL studies (see for an overview of learner behaviour for CALL design (Fischer 2007); for its use in SLA research (Chapelle 2007); and for its use for student modelling (Heift & Schulze 2007). Learner behaviour can be analysed to look for indications of when learning takes place (Collentine 2000; Mackey 2006; Chapelle 2007).

Studies that make use of computer mediated communication (CMC) have shown that they allow inspection of learner discourse and interaction (e.g. Sauro 2009), or learner reactions to CF (e.g. Diéguez-Bedmar & Pérez-Paredes 2012; Yilmaz 2012). More sophisticated uses of tracking learner behaviour are being implemented by adding eye tracking to determine when and where noticing takes place (e.g. Smith 2012). In CMC, the computer functions as a tool for interpersonal communication. A CALL system can also function as a tutor, responding to learner
output. Logging user-system interactions then informs us on how learners interact with a consistent tutor and can thereby highlight, and make comparable, the individual differences.

Besides detailed information on how much input and output the learner receives and produces, an important feature of learner behaviour is modified output as a result of CF. How learners respond to CF can be relevant for determining CF effectiveness (Lyster Ranta 1997; Nicholas et al. 2001). Indirect inferences about noticing can also be made through investigation of logged data in a CALL context (Hegelheimer & Chapelle 2000; Chapelle 2005), though it is difficult to assess noticing externally (Schmidt 1993).

Multiple studies by Heift (e.g. Heift 2002, 2004, 2010) investigated learner reactions to CF in a CALL context. A longitudinal study by Heift (2010) looked at learner uptake (repair of the error after CF). Learners of German (N=10) practiced for three semesters with a CALL system that gave two types of CF: error specific metalinguistic CF, and generic feedback in the form of metalinguistic clues. For advanced learners the error specific CF resulted in most uptake, whereas for lower level learners both types of CF resulted in equal uptake. Over the course of the semester, both types of CF were found to increase learner uptake. CF type thus related to amount of uptake (as found also in classroom studies), which differed according to learner characteristics. We cannot infer directly, however, that types of CF that elicit most uptake are the most effective for language learning. Heift writes in the conclusion that “the results make no claims on learning outcomes but instead focus on the learning process. For this reason, it still remains to be determined in what ways learner uptake ultimately contributes to L2 acquisition, in particular in a computer-aided language learning setting” (p. 212).

A CALL system is well suited to both perform analyses of learner behaviour and to measure learning outcomes. If these measurements are performed in a single CALL setup, then the findings of learner behaviour and learning outcome can be combined to provide insight into the role of modified output in language learning.

2.4. Research Questions

In the current study, we record both the product of L2 practice in proficiency pre- and post-tests, and the process of L2 practice through the logging of system-user interactions. This allows us to look in detail at two language learning conditions: L2 spoken output practice with immediate, automatic CF (CF condition) and spoken practice with the option to self-correct (NOCF condition). In this way, we want to answer the following research questions:

- **RQ.1:** Do the two treatment conditions have a different impact on learning Dutch verb second?
- **RQ.2:** To what extent does learner behaviour differ between the treatment conditions?
- **RQ.3:** What is the relation between learning Dutch verb second and learner behaviour, and does this relation differ between the treatment conditions?
3. Method

This study uses a CALL system that provides CF on spoken output through ASR. Learners work individually and complete grammar exercises in the form of interactive dialogues. This method allows for a detailed and individual approach that is highly consistent across multiple participants. We briefly describe it here, but refer to (Penning de Vries et al. 2014) for a more detailed description.

3.1. The GREET system

The GREET program is accessed online through a web browser. Learners log in and interact with the system through a graphical user interface. They are asked questions about film clips that they were shown, and are given words or segments of sentences graphically presented in ‘word blocks’ to form spoken responses. These responses are sent to the speech recognizer that maps the utterance to a representation in the language model, which contains all possible orders that can be created with the word blocks. It also outputs a confidence level that the recorded spoken utterance matches the selected utterance (for more details see Van Doremalen, Cucchiarini, & Strik 2010). If the confidence level is below a preset threshold, the system does not provide feedback but instead asks the learner to re-record. This is necessary to ensure against false accepts, or false rejects. Based on whether the answer sentence in the language model is labelled correct or incorrect, the system provides feedback to the learner. The type of feedback can be adapted to needs of the experiment (Section 3.4.2. describes the CF in this experiment). All speech recognition is performed on a central web server. This allows us to run simultaneous experiments, and to store all data centrally. The GREET system logs system-learner interactions to allow for insight into learner behaviour and system performance. Examples of logged data are: ASR output, learner spoken output, time spent reading instructions, time to answer, time per page, etc. Relevant for the current paper are the interactions with respect to each question, the number of attempts per question, and whether attempts were correct or incorrect.

3.2. Participants

We recruited 31 learners of Dutch (15 male, 16 female) at the language learning centre of our university from A1 and A2 level (CEFR) classes. They received 15 Euros for participating. The learners were on average 28 years old (SD=8.9, range 18-62), had lived in the Netherlands for 1 year (SD=1.1, 3 not specified), and had spent on average 4 months learning Dutch (SD=7.3, 5 not specified). They were all high educated, with most having a university degree. In total, 17 different L1s were represented. We grouped the languages according to word order in Table 1.
Prior instruction or knowledge of a target feature is assumed to have an effect on CF (Yilmaz 2012). Learners were controlled for their prior instruction and knowledge of the target structure by recruiting them at a specific CEFR proficiency level. It is necessary to specify learner proficiency on the basis of the pre-test scores on the target feature (see Section 4.3). The L1 is known to influence L2 learning (Ringbom 1992; Odlin 1989), and we expect transfer effects from languages that also have the target feature, such as German and Danish. However, the influence of transfer from the L1 can differ on account of other, additional L2s learned next to Dutch (Bohnacker 2006). This is discussed in Section 4.3.

### 3.3. Procedure

The experiment took place on two separate days at times selected by the participants. When they came in, they would first log in to the experiment website. On day 1, the participants filled in a short questionnaire about their background before starting with the proficiency tests, the grammaticality judgement test (GJT) and the discourse completion task (DCT) (see Section 3.5). Each test was preceded by click-through instruction web pages and three trial items. Both tests took 10-15 minutes to complete. After the proficiency tests, the participant moved on to 45 minutes of treatment (Section 3.4), which was also preceded by instruction pages. On day 2, one to six days later, the participants began with the treatment (45 minutes) and subsequently completed the proficiency post-tests, and finally filled in a post-test questionnaire (Section 3.6). In total, each session took 75-90 minutes.

---

8 During the treatment, learners were presented with a three-item questionnaire measuring their situated motivation every 15 minutes. This is not assumed to interfere with language proficiency practice, and is not discussed here. For a discussion on these questionnaires, see (Bodnar, Cucchiarini, Penning de Vries, Strik, van Hout (forthcoming))
3.4. Treatment

Participants watched short clips (30-45 s) of a video developed by an educational publisher about a man moving into a new apartment and meeting his neighbours (Nieuwe Buren, ‘new neighbours’, Put & Peekel 2001). After each video they were asked three to five questions about the clip. Responses were to be formed with word blocks (segments of the sentence) presented on the screen in random order. The learner would then press record and speak the sentence.

The treatment is designed in two parts. The first session contains 65 questions (25 target, 40 distracters); the second session contains 59 questions (19 targets, 40 distracters). Learners vary in their speed of practice, and some complete all questions of a session within the 45 minutes of practice. To control for time-on-task, learners were automatically restarted at the beginning to redo questions to complete 45 minutes of practice.

3.4.1. Target structure

The syntactic structure under investigation is inversion of subject and verb in declarative main clauses, as a result of verb second properties of Dutch. The learners are not informed that they are being trained or tested on this feature. Verb second (V2) entails that the finite verb appears in the second position regardless of the first constituent in the main clause. The subject precedes the finite verb when it is the first constituent. This order is inverted when another constituent precedes the finite verb. The subject then immediately follows the finite verb. This is illustrated in Table 2. The target structure occurs twice in the spoken video content (transcripts in Malmberg 1999). It does not occur in the question input of the treatment, only in the responses that the participants are required to construct.

<table>
<thead>
<tr>
<th>Target structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Subject-initial main clause</td>
</tr>
<tr>
<td>Jan</td>
</tr>
<tr>
<td>Jan</td>
</tr>
<tr>
<td>Subject</td>
</tr>
<tr>
<td>(2) Adverbial-initial sentence: Postverbal subject</td>
</tr>
<tr>
<td>Morgen</td>
</tr>
<tr>
<td>Tomorrow</td>
</tr>
<tr>
<td>Adverbial</td>
</tr>
</tbody>
</table>

Table 2. An illustration of the target structure in our study: the subject-verb order inverts when the first constituent of the sentence is not the subject, to comply with verb second.
V2 is an appropriate feature to study, as the acquisition of subject-verb inversion is problematic for L2 learners of Dutch (Jordens 1988). A crucial argument is that V2 does not affect the meaning of the sentence, and is therefore likely to be non-salient in the input and unlikely to prompt CF in meaning-based communication. Grammatical items that have no direct relation to meaning take the longest for L2 learners to acquire (R. Ellis 1994; Gass & Selinker 2008).

3.4.2. Corrective Feedback in our CALL system

Figure 1. Flowchart of how the GREET system responded to spoken learner attempts in the two experimental conditions.

Figure 1 illustrates the differences between the two conditions. In the CF group, the system explicitly told the learners whether their utterance was correct or incorrect. For a correct answer, the learner was shown a green check mark and moved to the next question. For an incorrect answer, the learner was shown a red screen saying ‘That is incorrect. Please try again’ and they were returned to the same question. After a second incorrect attempt, the system started to give hints. One of the word blocks is placed in its appropriate place in the answer sentence. To prevent
frustration (due to either learner error, or ASR error) the system makes a skip-button available after four wrong tries. If a learner practices the same sentence ten times, the system shows a message explaining that there may be something wrong, and suggests that they skip this question (non-optional). In the NOCF group, after each recording by the learner, the system gives them the following message in a white pop-up screen: “You can only save one answer. Do you wish to save this one and move to the next question?” They then have the options: (1) retry, or (2) move to the next question.

3.5. Proficiency tests

Pre- and post-tests measured changes in proficiency. We employed a timed grammaticality judgment test (GJT), which involves receptive (written) knowledge, and a discourse completion task (DCT) for productive (spoken) knowledge. The focus of analysis is on the DCT, as we target spoken competence, but the GJT is included for cross-task validation (cf. Ellis 2005; Norris & Ortega 2003) and informs us on task-specific improvement through a comparison of distracter and target items.

3.5.1. Discourse completion task (DCT)

The design of the DCT was based on Van de Craats (2009), but was implemented in a computer environment. In the DCT, the participants had to record a spoken sentence. Participants were given the first part of a sentence that they had to complete. They saw a picture pertaining to the sentence, with below the picture a noun, and an unconjugated verb, which they could use in their response (see Example 1). This was to ensure that the learners would not have problems completing the sentence as a result of lexical retrieval. When the learners had recorded a response, they could click ‘next’ to be presented the next question. The DCT was counterbalanced in two sets of 32 questions. The two sets were matched so that each of the 32 questions in set A were matched with a question with similar grammatical structure and lexical content in set B. Learners were alternatively given set A or B as pre-test and accordingly set B or A as the post-test. Each set had 16 target questions and 16 distracter questions. Distracter questions pertained to features that were known to be common errors in L2 Dutch. The order of the DCT questions was randomized for each participant. For each item there was a time limit of 30 seconds. If the participant did not record an answer within that time, the response was scored as incorrect.
Example 1. DCT target item

[illustration to the right of the question: photo of sandwich]

Question:
Wat doet Kim om twaalf uur?
‘What does Kim do at twelve o’clock?’
"Om twaalf uur…”
‘At twelve o’clock …’

boterham + eten
‘sandwich + to eat’

3.5.2. Timed grammaticality judgment test (GJT)

In the GJT, participants judge whether the sentence on the screen is correct in Dutch. They respond by clicking with the mouse. The test was counterbalanced, in two sets of 40 test items. Each set had 20 target items, and 20 distracter items, of which half the items were grammatical. The order of the items was randomized for each participant. For each item there was a time limit of 12 seconds (set in pilot experiments). Responses outside that time are scored as incorrect. The timed GJT is a receptive reading task, but it was timed to give the learners a sense of pressure in order to tap knowledge that is assumed to be associated with implicit knowledge (see the psychometric study by Ellis 2005), which underlies speaking proficiency.

3.6. Learner evaluation questionnaire

At the end of the second session participants completed a questionnaire to evaluate the system and their learning experience. There were 19 questions for the CF condition, and 17 questions for the NOCF condition (see Appendix H for the questions). These questions were answered on a semantic differential 5-point scale. Nine questions about usability, enjoyment, and perceived effectiveness of the program were shared for CF and NOCF. The other questions concerned preferences for modality and CF, and control questions. In addition, two open questions asked about preferences for feedback, and there was room for general comments.

3.7. Logged data of learner behaviour

The system logs learner-system interactions, and stores the recorded utterances. This way we have access to information about time, and duration (e.g. log-in time, and time spent on each page), and about the specific interactions related to the practice exercise, (e.g. the number of attempts per question). System actions are also logged (e.g. the recognition result of the learner's utterance, and the CF presented).
4. Results

We first report on the effect of the two treatment conditions as measured by the proficiency tests (4.1, 4.2), followed by an exploration of other differences between the conditions (language background (4.3), post-test questionnaire (4.4)). Then we proceed with an analysis of logged data of learner behaviour (4.5).

4.1. Discourse completion task

The DCT item reliability statistics indicate that it is a reliable test (Cronbach’s alpha for the 32 target items=.960; scores above 0.7 are generally considered reliable).

Figure 2 shows the individual trajectories of the DCT scores split out by treatment condition. The DCT pre-test means are (M(CF)=.39, SD=.34, range: .00-.94; M(NOCF)=.30, SD=.32, range: .00-.94). There is no difference between the groups at pre-test (F(1,30)=.577, p=.453), but there is large variation in the pre-test scores from zero to almost full competence with V2.

We recruited learners at a lower proficiency level than our previous study (Penning de Vries et al. 2014) namely A1-A2, instead of A2-B2 (CEFR level), to reduce the number of learners performing at ceiling level. The CEFR levels line up with the V2 proficiency mean scores: (M (A1) = .29, SD= .30; M (A2) = .46, SD = .37). However, within both CEFR levels the individual V2 proficiency scores still range from zero knowledge to full competence (range: .00 -.94).

Pre-test proficiency is assumed to influence and effectiveness of the CF and the NOCF condition - obviously so in the case of ceiling effects. It is expected to influence both proficiency gain and learner/learning behaviour during practice. To account for ceiling proficiency effects, we proceed to analyse the data in two sets: (1) the complete group of 31 learners, and 2) a trimmed group of 24 learners, excluding those with a DCT pre-test score over 0.6. This cut-off point for the trimmed group is based on: combined observations in the log data (see Section 4.5), learner characteristics (see 3.2 and 4.3), and V2 proficiency scores as measured on the pre-tests. Also for the trimmed group, there is no difference in the pre-test scores between the conditions (F(1,23) = .579, p=.455).
Figure 2. Line graphs of the individual scores (proportion correct) on the DCT pre-test and post-test, split out by NOCF and CF condition.

Throughout this paper we will present the data in these two groups of successful and unsuccessful learners. The descriptive data for the trimmed and complete CF and NOCF groups are given in Table 3.
In Table 4 the results of DCT performance are analysed in a repeated measures ANOVA (within-subjects factor time, and between-subjects factor condition).

<table>
<thead>
<tr>
<th>DCT mean scores</th>
<th>Complete group</th>
<th>Trimmed group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre M (SD)</td>
<td>post M (SD)</td>
</tr>
<tr>
<td>CF</td>
<td>.39 (.344)</td>
<td>.53 (.295)</td>
</tr>
<tr>
<td>NOCF</td>
<td>.30 (.318)</td>
<td>.37 (.347)</td>
</tr>
</tbody>
</table>

Table 3. Means and SDs of proportion correct on the DCT proficiency tests, CF vs. NOCF for the complete (CF N=16, NOCF N=15) and trimmed group (CF N=12, NOCF N=12).

<table>
<thead>
<tr>
<th>ANOVA of DCT scores</th>
<th>Complete group (N=31)</th>
<th>Trimmed group (N=24)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>F(1,29)=13.912, p=.001, η²_p=.324</td>
<td>F(1,22)=2.429, p=.000, η²_p=.481</td>
</tr>
<tr>
<td>Condition</td>
<td>F(1,29)=1.216, p=.279, η²_p=.040</td>
<td>F(1,22)=2.042, p=.167, η²_p=.085</td>
</tr>
<tr>
<td>Time by Condition</td>
<td>F(1,29)=1.515, p=.228, η²_p=.050</td>
<td>F(1,22)=4.352, p=.049, η²_p=.165</td>
</tr>
</tbody>
</table>

Table 4. Results of the DCT in a 2x2 mixed factorial ANOVA for the complete and the trimmed group.

Table 4 shows that spoken practice with GREET is beneficial for both conditions, as there is a significant effect of time. To determine whether CF is more effective than the NOCF condition, we look at the interaction of condition by time. This is non-significant for the complete group, but it is significant for the trimmed group (this significant interaction effect is also found when the cut-off boundary of the trimmed group is lowered, but the analysis loses power due to fewer participants when we lower the cut-off point too much).

We proceeded to analyse the conditions separately, in a repeated measures ANOVA. Here we find that for the CF condition there is a significant effect of time (F(1,15)=10.041, p=.006, η²=.401; trimmed F(1,11)=18.470, p=.001, η²=.627), and for the NOCF condition the effect of time is not significant (F(1,15)=4.232, p=.059, η²=.232 ); trimmed: F(1,11)=3.617, p=.084, η²=.247). The effect size is found to be larger for the CF condition.

Results of the DCT show that practice in the CF condition is more beneficial than the NOCF condition for learning Dutch V2, and that learning, and the effect of
CF, in the treatment are related to proficiency: low and medium proficient learners benefit more from practice with CF.

4.2. Grammaticality judgement test

The GJT is found to be reliable (Target items is 40, Cronbach’s Alpha = .752). For the fillers, the reliability is lower (GJT fillers is 40, Cronbach’s Alpha = .559). In the filler items different types of common L2 errors are featured, which seem to have led to lower reliability. There is no difference between the groups in the GJT pre-test scores. (For the target items: complete F(1, 30) = 1.237, p = .275; trimmed: F(1, 23) = 1.817, p = .191. For the filler items: F(1, 30) = 3.067, p = .090; trimmed: F(1, 23) = 2.984, p = .098). The GJT scores are correlated with the DCT scores (pre-test scores: r = .585, p = .001; trimmed group: r = .225, p = .291; post-test scores: r = .621, p = .000; trimmed group: r = .553, p = .005). In interpreting this correlation, we need to take into account that the GJT is a receptive, written, forced choice test.9 The descriptive statistics of the GJT are given in Table 5. In an ANOVA we test for significant effects on GJT performance. These results are given in Table 6.

<table>
<thead>
<tr>
<th></th>
<th>Complete group</th>
<th>Trimmer group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRE M (SD)</td>
<td>POST M (SD)</td>
</tr>
<tr>
<td>Targets CF</td>
<td>.56 (.15)</td>
<td>.65 (.17)</td>
</tr>
<tr>
<td>Targets NOCF</td>
<td>.49 (.17)</td>
<td>.55 (.20)</td>
</tr>
<tr>
<td>Fillers CF</td>
<td>.65 (.09)</td>
<td>.64 (.13)</td>
</tr>
<tr>
<td>Fillers NOCF</td>
<td>.59 (.10)</td>
<td>.63 (.13)</td>
</tr>
</tbody>
</table>

Table 5. Means and SDs for the target and filler items of the GJT. CF vs. NOCF, for the complete (CF N=16, NOCF N=15) and trimmed group (CF N=12, NOCF N=12).

9 Because the GJT is a forced choice test, guessing and/or mistakes influence the score. Guessing may cause learners who do not know the answer to still provide a correct response.
The Role of Automatic CF in Learning L2 Grammar

The results in Table 6 show that there is a significant effect of time for the target items. This means that the participants improved on the target items. For the target items, there is no interaction of time by condition, thus the CF and NOCF group improved equally on targets as a result of training. For the trimmed group, improvement on targets is (just) non-significant. Also there is no interaction of time by condition, showing that for the trimmed group there is no difference in learning effects on the GJT between the conditions.

The scores on filler items show no significant effects as a result of treatment, i.e. there is no improvement on filler items. This shows that the learning effect found on targets is not attributable to test familiarity. There is also no interaction of time by condition for the fillers.

### 4.3. Language background

In an analysis of L1 and DCT performance, we find that L1 German participants perform at ceiling level. Apparently German inversion of V2 in the main clause transfers to this Dutch learning exercise, and is not (significantly) influenced by their other L2, English (cf. Bohnacker 2006). Otherwise we did not find evidence that a particular L1 or L1-word order had an advantage or disadvantage.

Besides knowledge of L2 Dutch, 29 of 31 participants indicated knowledge of another L2. Mostly this L2 was English (28), and probably at a high level, since the...
experiment was conducted in an academic environment. We do not expect an influence of English (SVO, non-V2), and pre-test scores show no evidence of an influence as several learners of English scored zero on the DCT. There were six participants with German L2 (SVO, V2). Four of these participants scored over 60% correct (68%, 80%, 94% and 94%) on the DCT pre-test, which excludes them from the trimmed group. The other two participants with L2 German improved quite dramatically on the post-test: one was a CF participant, the other a NOCF participant, with Hungarian and Russian L1, respectively.

For L1 knowledge, we can assume transfer effects. For transfer of L2 knowledge we need to take into account their level of proficiency in that L2. If a learner does not master V2 in the L2, transfer is obviously not possible. A participant who indicated L2 knowledge of Danish (SVO, V2), scored 6% correct on the pre- and post-test. In this case there is no reason to assume a level of knowledge of Danish V2 to allow for transfer effects. As a result we do not categorically exclude all learners with German or Danish as an L2.

4.4. Post-test evaluation

In a post-test questionnaire that measured the learners’ evaluation of the system, nine questions were shared in the CF and NOCF questionnaire that directly relate to GREET practice. The reliability score for these nine items is acceptable (Cronbach’s Alpha=.744) so we take the mean score of the shared items to make a between group comparison. The means scores are given in Table 7.

<table>
<thead>
<tr>
<th></th>
<th>Complete group</th>
<th>Trimmed group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (SD); range</td>
<td>mean (SD); range</td>
</tr>
<tr>
<td>CF</td>
<td>4.10 (.43); 3.3 - 4.7</td>
<td>4.03 (.44); 3.3 - 4.7</td>
</tr>
<tr>
<td>NOCF</td>
<td>3.76 (.64); 2.8-5.0</td>
<td>3.51 (.42); 2.8-4.1</td>
</tr>
</tbody>
</table>

Table 7. The means and SDs for the post-test questionnaire, CF vs. NOCF, in the complete (CF N=16, NOCF N=15) and the trimmed group (CF N=12, NOCF N=12).

In Table 7 the mean for the CF seems higher in both complete and trimmed group. However, statistical analysis does not show a significant difference between conditions for the complete group (F(1,29)=3.058, p=.091), but the difference is significant for the trimmed group (F(1,22)=8.577, p=.008). This can be explained by looking at Figure 3: lower proficient learners in the NOCF condition evaluated the system less positively, whereas the rating in the CF condition is relatively constant, and unrelated to pre-test proficiency. This is confirmed in a correlation analysis: pre-
test proficiency correlates with mean evaluation score for the NOCF group (r=.760, p=.001), but not for the CF group (r=.262, p=.326); correlations for the trimmed group are not significant. In this way the findings on the DCT, where the trimmed group benefited more from CF, can be related to the learner’s perceived efficiency and enjoyment of practice: the group that learned more from treatment, is more positive.

For the individual items on the questionnaire, the CF group scored two items significantly higher: Question 1 in the trimmed group (F(1,22)=6.401, p=.019) and Question 6 in the complete and trimmed group (F(1,29)=5.194, p=.030 and F(1,22)=5.852, p=.024). Question 1 (“practice with GREET is a (bad – good) way to improve my Dutch”) relates directly to whether the participants thought the program improved their Dutch. Question 6 (“The videos are (difficult – easy) to understand”) indicates that that the CF condition assisted the learners in understanding the video. The questions in the exercise are all about the videos, and the CF may have required them to practice questions more often and thus giving more time for reflection; or more indirectly, the CF may have given the learners more confidence that they understood the video, through the positive (green check mark) feedback.

Figure 3. Scatterplot showing the mean evaluation score per participant, plotted against their DCT pre-test score.
4.5. Learner behaviour

This section presents data of learner behaviour during individual (unsupervised) practice. In particular we look at how practice differs per treatment condition, and how learner behaviour relates to V2 proficiency. The following sections present data on the number of questions practiced (4.5.1), the number of attempts to answer the questions and how that changed over time (4.5.2), and the number of times learners produced a correct answer and how the accuracy of the attempts changed over time (4.5.3).

4.5.1. Questions practiced

The treatment session contains 124 questions, of which 44 are target questions. If learners complete all questions, they restart and continue until 45 minutes of practice time is up (referred to as 'looping'). Table 8 gives the number of questions practiced during 90 minutes of practice. The mean for the complete group exceeds the maximum number of unique (target) questions, showing that several learners looped. For the trimmed group the number of questions practiced are close to the maximum number of unique questions, this indicates that the content and the practice time is more appropriate for the (proficiency level of) trimmed group.

<table>
<thead>
<tr>
<th>Questions practiced</th>
<th>Complete group</th>
<th>Trimmed group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Mean (SD)</td>
<td>Targets Mean (SD); range</td>
</tr>
<tr>
<td>CF</td>
<td>142.94 (42.95)</td>
<td>47.56 (15.40); 27-81</td>
</tr>
<tr>
<td>NOCF</td>
<td>141.4 (27.55)</td>
<td>47.13 (9.58); 32-66</td>
</tr>
</tbody>
</table>

*Table 8. Means and SDs of the number of questions (all and target only) in 90 minutes of GREET treatment, CF vs. NOCF, for the complete and the trimmed group.*

Table 8 shows that the number of target questions practiced is similar for both groups (complete: F(1,29)=.009, p=.927; trimmed: F(1,22)=.376, p=.546)). Figure 4 plots the mean number of questions practiced against DCT pre-test proficiency score.
The Role of Automatic CF in Learning L2 Grammar

The pre-test scores are correlated to the number of questions practiced (CF(complete): $r=.830$, $p=.000$; CF(trimmed): $r=.687$, $p=.013$; NOCF(complete): $r=.706$, $p=.003$, NOCF(trimmed): $r=.126$, $p=.696$). This is also seen in Figure 4. The stronger correlation found for the CF group is likely to be because the system intervenes in the learners’ activities, and does so more when more errors are made.

4.5.2. Attempts per Question

A learner can make multiple attempts at answering a question in the treatment. For the NOCF group this is self-regulated and is potentially infinite; for the CF group, the number of attempts per question (ApQ) is regulated by the system: if it assesses the response as incorrect, the learner is prompted to try again.

In the CF group, learners who make errors are (a) made aware of the error and prompted to repair the utterance, and (b) after a second incorrect attempt they are given a hint to repair the utterance. When the system gives a hint, it fills the first open position in the sentence with the appropriate word block. This continues...
sequentially until all blocks have been placed. For 28 (64%) of the target questions, the first hint will fill the V2 position (the block containing the finite verb). For the remaining 16 (36%), it is the second hint that places the verb in V2 position.

This means that for ApQ=1, the learner gave the correct answer in the first try; for ApQ=2, the learner successfully repaired an error after explicit information that their answer was incorrect; for ApQ> 2 the learner repaired an error after explicit information that the answer was incorrect, and with hints from the system.

Complete group | Trimmed group
--- | ---
Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD)

<table>
<thead>
<tr>
<th></th>
<th>Attempts</th>
<th>ApQ</th>
<th>Attempts</th>
<th>ApQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>84.81 (21.65)</td>
<td>1.93 (.724)</td>
<td>84.50 (22.34)</td>
<td>2.14 (.715)</td>
</tr>
<tr>
<td>NOCF</td>
<td>55.40 (18.84)</td>
<td>1.17 (.311)</td>
<td>50.83 (17.82)</td>
<td>1.16 (.336)</td>
</tr>
</tbody>
</table>

Table 9. Means and SDs for the number of attempts, and attempts per question (ApQ) in 90 minutes of GREET treatment, CF vs. NOCF, for the complete and the trimmed group.

The number of attempts and ApQ is given in Table 9. Applying ANOVAs shows that the difference between the CF and the NOCF for the number of attempts is significant (F(1,30)=16.183, p=.000); trimmed group F(1,23)=16.652, p=.000). The difference for ApQ is also significant (F(1,30)=13.817, p=.001; trimmed group: F(1,23)=18.254, p=.000). The difference between the conditions is most pronounced in the trimmed group. As seen in Figure 5, the difference in ApQ is largest with lower proficient learners.
The Role of Automatic CF in Learning L2 Grammar

Figure 5. Scatterplot of the mean number of attempts per question in 90 minutes of treatment, against the DCT pre-test proficiency score. In the CF condition learners are given hints after the second attempt.

For the NOCF group the number of ApQ is approximately one, regardless of V2 proficiency. Learners do not seem to retry when they make an error. As a result, the number of ApQ for NOCF is not correlated with V2 proficiency level (r = -0.063, p = .824). For the CF group, there is an inverse correlation of ApQ with V2 proficiency (r = -0.653, p = .006): lower proficient learners practice more ApQ than the higher proficient learners. This is expected, as GREET requires learners in the CF group who make an error to retry until they provide a correct answer.

10 One NOCF participant practices more (2.13 ApQ) than all other NOCF participants (M= 1.1 ApQ). This participant is low proficient, but seems to be very motivated to practice speaking. The learner, however, does not have knowledge of V2 and his retries do not address V2, even though he seems to focus on word order. He does not improve on the post-test.
11 One CF participant has a remarkably high number of ApQ. On inspection we find that he has halting and slow pronunciation, but ASR recognition and CF is accurate. He seems to be having problems with lexis as well. We inspected the number of ApQ over time, and find that per 15 minute interval his ApQ changes from 4.5, 5, 4, 3, 3.7, 2.8. His ApQ decreases over time, indicating that his performance on the task improves, coinciding with improvement on the DCT. He uses the option of skipping a question 6 times (out of 29 possibilities).
For ApQ, it is relevant to inspect changes over time, to see if learners require fewer attempts to answer a question as training proceeds. In Figure 6 the mean number of ApQ is given per 45-minute practice session, to show the changes between the sessions.

**Figure 6. Bar graphs with standard errors (± 2SE) of attempts per question (ApQ) per 45-minute session, for the complete group and the trimmed group.**

Figure 5 shows that the ApQ for the trimmed CF group drops from 2.31 (SD=.87) to 1.97 (SD=.58). This means that the participants are moving from receiving hints from the system (over 2 ApQ) to more often correcting after explicit CF, without a
The Role of Automatic CF in Learning L2 Grammar

hint (under 2 ApQ). In an ANOVA we find that the decrease of ApQ is significant (Effect of time: $F(1,29)=5.690$, $p=.024$, $\eta^2_p=.164$; trimmed $F(1,22)=6.174$, $p=.021$, $\eta^2_p=.219$), and we also find a difference between the groups over time (Interaction of time by condition: $F(1,29)=7.490$, $p=.010$, $\eta^2_p=.205$; trimmed $F(1,22)=8.247$, $p=.009$, $\eta^2_p=.273$). In Figure 6 we can see that the difference between the groups is that the NOCF condition is stable at 1 ApQ, while the CF condition decreases in ApQ.

The NOCF group improves their V2 proficiency—though to a lesser extent—without retrying on questions. This behaviour does not change over time. If learners noticed their error(s), then they very infrequently tried to self-correct the error.

4.5.3. Number of OKs

The number of OKs indicates how many times participants provide a correct answer. In the NOCF condition, production of the correct utterance relies on the learners’ knowledge, and on self-monitoring of their responses. The ASR does process the learner’s output, but the assessment is not shown to the learner. In the CF condition the learners have to retry until they produce a correct sentence (i.e. see a green check mark on the screen indicating an “OK”), before being allowed to proceed to the next question. Table 10 shows the number of OKs: the number of times the learner produced a target sentence correctly (i.e. applied V2 inversion correctly).

<table>
<thead>
<tr>
<th>Number of OKs</th>
<th>Complete Mean (SD)</th>
<th>Trimmed Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>44.75 (16.72)</td>
<td>38.08 (12.09)</td>
</tr>
<tr>
<td>NOCF</td>
<td>19.07 (2.43)</td>
<td>10.33 (8.86)</td>
</tr>
</tbody>
</table>

Table 10. Means and SDs of number of OKs in 90 minutes of GREET treatment, CF vs. NOCF, for the complete and the trimmed group.

The number of OKs for the NOCF condition is significantly lower (complete: $F(1,29)=14.757$, $p=.001$; trimmed: $F(1,22)=41.127$, $p=.000$). This has two reasons: (1) in the CF condition, the learners must retry until they provide a correct answer to move on, whereas NOCF learners can proceed without providing a correct answer, and (2) the NOCF learners do not receive feedback on their performance and sometimes make errors in recording (e.g. inadvertently truncating the audio, or restarting an utterance). The trimmed group has a lower number of OKs which is self-evident as the ceiling learners were removed.

Though this is a confound, the fact that the high proficient learners are close together in number of OKs suggests that this is not a major confounding factor.
Figure 7. Scatterplot showing the number of correctly formed sentences (OKs) during 90 minutes of treatment per participant, against pre-test proficiency.

Figure 7 shows that for the CF and NOCF condition, the number of OKs correlates with the initial V2 proficiency level (CF: \( r = .889, \ p = .000 \), NOCF: \( r = .884, \ p = .000 \)). In addition, we see that the difference between the CF and NOCF group in number of OKs becomes smaller as pre-test proficiency increases.

To investigate learning behaviour, we look at changes over time. Because the CF group is automatically led to an OK, we look instead at the ratio of OK per attempt. Figure 8 compares the participants’ performance in the first and second session.
Figure 8. Bar graphs with standard errors (± 2SE) showing the ratio of OKs per attempt per 45-minute session for the complete and the trimmed group.

Figure 8 shows the learner’s accuracy per attempt. If proficiency improves over time, the ratio should become higher: i.e. an attempt is more often a grammatical (correct) response. In an ANOVA we find that the ratio of OKs changes over time (Effect of time: complete $F(1,29)=4.294$, $p=.047$, $\eta^2_p=.129$; trimmed $F(1,22)=5.985$, $p=.029$, $\eta^2_p=.218$).
There is no difference between the groups in the change over time (Interaction time by condition: complete $F(1,29)=1.533$, $p=.226$, $\eta^2_p=.050$; trimmed $F(1,22)=.543$, $p=.469$, $\eta^2_p=.024$). As a result, we find that for both conditions the ratio of OKs per attempt improves.

On closer inspection of the NOCF group, we find that the ratio of OK per attempt decreases for the high V2 proficient learners. We find that accuracy drops in the last 15 minutes of practice and that they are repeating questions. A decrease of concentration and/or motivation is likely to be the cause of the lower accuracy.

5. Discussion and future perspectives

In this section we discuss the results of the study per research question and then proceed to a general discussion of their implications and suggest future perspectives.

5.1. Do the two treatment conditions have a different impact on learning Dutch verb second?

Learners practiced for 90 minutes with our CALL system, producing spoken output in a dialogue-based task that targeted V2. Learners in both the CF condition and the NOCF condition improved their score on a spoken production task (DCT) and on a speeded receptive reading task (GJT). For the group overall we do not find a difference in learning between the CF and the NOCF group. When higher V2 proficient and ceiling learners are excluded, the CF condition outperforms the NOCF condition. As a result, we find that providing CF facilitates language learning more than spoken production alone for learners with low to medium knowledge (0-60%) of the target structure.

In the CF condition, negative CF is only provided in case of a learner error. If the learner is high-proficient, only few instances of negative CF will be provided. In a way, therefore, the behaviour of the system (and of the learner) in the CF condition will start to resemble that of the NOCF condition more closely, as the learner's proficiency is higher and fewer negative CF messages are seen. On the other hand, the learners in the CF condition do receive positive CF (OKs). The effect of the positive CF may be smaller in high-proficient learners as their confidence is likely to be higher, and they expect their response to be correct (Hattie & Timperley 2007). In addition, higher proficient learners are more capable of noticing their errors (Hanaoka 2007; Ellis 2000), thus further reducing the need for CF. Taken together, for learning V2 (our target feature) learners benefit more from spoken practice with CF than from spoken practice without CF in the initial stages of language learning, but that benefit decreases as L2 proficiency increases.

The improvement found in the separate conditions may be the result of different processes. In the CF condition, learners produce output containing V2 and are given explicit information on whether the structure was used correctly, and after 2-3 incorrect attempts they are given input regarding the target structure. Thus
learners are (a) prompted to retry incorrect answers, and pushed to produce output: which is beneficial for L2 learning according to the Output Hypothesis (cf. Swain 1985; De Bot 1996); (b) assisted in producing the correct utterance (proceduralization) in a meaningful and skill-specific context (cf. DeKeyser 2007), (c) assisted in noticing the error (negative evidence) (cf. Schmidt 1995; Kachtava & Ammar 2013), (d) and are provided with input regarding the correct form (positive evidence) (cf. VanPatten 2004), which are all beneficial circumstances for L2 learning.

In the NOCF condition, learning relied on the learner’s own ability to notice an error and to produce a grammatically correct sentence during practice. The NOCF learners were likely to be focused on linguistic form, as a result of the task they were practicing and because they were told that their practice was automatically evaluated by the ASR and scored. Compared to practice in the CF condition, learners in the NOCF condition are also (a) pushed to produced output since the task requires them to record at least one attempt in order to proceed. Potentially they (b) produce the correct form, and (c) notice L2 errors, but (b, c) are learner dependent: the learners were stimulated to produce grammatically correct sentences by promising them a score at the end of practice, and perhaps intrinsically by the learner’s belief of efficient practicing. However, efficient practice with respect to (b,c) then depends on their proficiency level and whether it was sufficient to be able to produce a correct sentence, and/or notice an error. Learners in the NOCF condition did not (d) receive input. Since the learners in the NOCF condition were seen to improve, we find that producing the sentences is also beneficial without CF. Sanz and Morgan Short (2004) also found that exposing learners to task-essential practice is sufficient to promote acquisition, though they find that pre-test explanation and explicit CF during practice did not determine learning outcome, whereas we find an effect for CF. Perhaps the improvement can be explained because the NOCF group spends more time per attempt; the reliance on self-monitoring (and the absence of CF) may cause the NOCF learners to be more cognitively involved than the CF learners in trying to produce a correct utterance, which is beneficial for language learning (cf. Robinson 2003).

The low to medium proficient learners in NOCF condition seem to improve, but this varies per participant: some do not improve on the post-test. These learners may have to address multiple aspects of sentence production, besides correctly applying V2. Since they have to self-correct and have not been told what the target structure is, they may direct their attention to (i.e. notice) other issues in their response (e.g. pronunciation, lexis) (cf. Skehan 1998: 112). Also, learners with zero knowledge of the target structure cannot self-correct (e.g. one NOCF learner frequently retries, but addresses mostly pronunciation, and is not aware of his V2 errors). By comparison, the CF condition implicitly focuses learners’ attention on word order and V2, by signalling errors and having participants retry until they produce a correct V2 sentence. Lyster et al. (2013: 13-14) report on disagreement among researchers on the effectiveness and necessity of CF for learning new linguistic forms, or whether it is only relevant for consolidation of partially acquired
knowledge: in other words, whether CF is helpful for learners starting at zero level. In our system, we find that these learners do improve with CF that explicitly signalled the presence of an error and gradually provided hints.

Between the groups, there is a difference in the learners’ evaluation of practice with the system. For low to medium proficiency participants, the CF group was more positive about practice than the NOCF group. The attitude towards practice in the CF group was unrelated to V2 proficiency level, but in the NOCF group lower proficient learners were less positive about practice than higher proficient learners. Taken together with the proficiency test results, this suggests that the effectiveness of the practice condition influences the learner’s evaluation of the system, and that the presence of CF is felt to be more important for the low proficient learners. The favourable attitude regarding CF on oral correction is in line with findings by Schulz (1996), that, contrary to teacher’s belief, learners were rather positive about oral error correction; but it could also reflect the more pragmatic attitude reported in Jean and Simard (2011), that learners think that error correction and grammar is effective, even though they do not enjoy it.

Another learner characteristic we inspected was language knowledge. We found indications of transfer of the V2 target structure from L1 and L2 German (cf. Ringbom 1992). We did not find other indications of influences of the L1 among the 17 different L1s that were present in our participant sample.

5.2. To what extent does learner behaviour differ between the treatment conditions?

We find that the learners in both conditions practice a similar number of questions, but that the CF group produces more attempts at answering these questions, and that they produce more grammatically correct V2 sentences. The differences between the groups are largest for the low to medium V2 proficient learners. The CF condition, because it pushes learners to produce more (grammatical) output, is likely to be more beneficial for language learning, as skill-specific practice is necessary for proceduralization: spoken production of correct instances of V2 structure improves speaking proficiency (cf. DeKeyser 2007).

In the CF condition, participants were prompted to retry the question in case of an error, until they had successfully repaired the error. Thus, with each error, the learner had an opportunity to notice the error, and to produce the correct form. Even though the CF was generic in only stating the presence of an error, it allowed learners to test their hypotheses of L2 grammar (a reason why output production facilitates SLA according to the Output Hypothesis (Swain 1985)). In addition, the CF condition provided hints after repeated errors, thereby providing input regarding the target form. As also learners who were zero V2 proficient on the pre-test improved, these learners must have picked up information regarding V2 from the CF.
The Role of Automatic CF in Learning L2 Grammar

In addition, the CF prompted modified output from the learners, leading to repair of the error. Interactions when learners correct errors in their linguistic output may suggest noticing (cf. Chapelle 2005). The role of repaired utterances may be significant for language learning because they provide learners with opportunities for learners to proceduralize target language knowledge (Ellis et al. 2001: 282). The NOCF group infrequently retried questions, and repaired an error very infrequently (successful repair was found for 23 out of 708 questions practiced (=3%)). Van der Linden (1993) also found that if learners do not receive CF, they generally do not try to correct themselves. The role of repair of errors following CF (uptake) for language learning has been a source of discussion (Lyster et al. 2013); though several types of CF have been shown to elicit more uptake (e.g. Lyster & Ranta 1997, Heift 2010), it has not yet been shown convincingly that uptake results in better learning results (cf. Loewen 2004). In this study we find that the CF condition generates more uptake from the learners, and that the CF group improves more on the proficiency post-test than the NOCF group. However, since the NOCF group does improve, we also find that uptake is not necessary for L2 learning.

5.3. What is the relation between learning Dutch verb second and learner behaviour and does this relation differ between the treatment conditions?

In both the CF and the NOCF condition, we found that the V2 proficiency level of the learners correlated with the number of questions practiced, and the number of OKs. In other words, the lower proficient learners practice with fewer instances of V2 (questions) and produce fewer grammatically correct sentences (OKs). For the learners in the CF condition, the number of ApQ was correlated with proficiency: lower proficient learners practiced more ApQ. Learners in the NOCF condition very infrequently retried questions, regardless of their proficiency level. Learners have been shown to vary largely in how they interact with a CALL program, which is in part related to their proficiency (Hegelheimer & Tower 2004; Heift & Rimrott 2012). How learners interact with a CALL program is likely to determine the effectiveness of the program. Since we find that practice was more effective in the CF condition for the low to medium V2 proficient learners, it is important to see how learner behaviour relates to learner proficiency.

Lower V2 proficient learners practice fewer questions than higher proficient learners, but this is similar in the CF and the NOCF group. Thus the practice (and input) with unique questions is comparable for both groups and does not explain the difference in the learning outcome, though it obviously is important for language learning (cf. VanPatten 2004, Sanz & Morgan-Short 2004).

The relation of V2 proficiency with ApQ differs per condition. In the CF condition, low proficient learners practiced the most ApQ, which means that these learners appropriately receive the most CF. In the NOCF condition, ApQ is not correlated with proficiency level. In general, regardless of proficiency level, NOCF
learners move to the next question after one attempt, without retrying. Here, the CF condition is likely to have been more beneficial for low proficient learners, as it patiently points out the error, until the learner corrects it (with or without input in the form of a hint), thereby increasing the chances of noticing of the target structure. Moreover, it causes the learners to repair their error and ultimately produce a grammatically correct sentence (an OK).

The number of OKs correlates with proficiency level in both conditions, but the NOCF group produces a lower number of OKs overall. The difference between the conditions is largest for the low proficient learners. We may assume that a minimum number of OKs is necessary for improvement, as learners in the NOCF condition with the lowest number of OKs did not improve on the post-test. These learners were found to be zero-level learners. By comparison, CF learners who were zero-level proficient improved, and produced approximately twice as many OKs. This suggests a minimum number of (spoken) productions of the correct form is necessary for improvement. Also relevant is that learners in the CF condition received confirmation that their utterance was correct, in the form of a green check mark. Here, the CF condition is likely to build their confidence, whereas the NOCF condition is neutral in this respect (however, we can imagine learners becoming more aware of the structure they are uncertain about in subsequent output and input, and thus increase chances of noticing).

The improvement of the learners' proficiency can be seen in learner behaviour, as changes over time in the treatment. The number of ApQ decreases for the low to medium proficient learners in the CF group. This means that they need fewer instances of CF to repair their error as treatment progresses. Importantly, they were seen to rely less on hints. However, the NOCF participants generally do not self-correct errors, and this does not change over time. Improvement of proficiency in the NOCF condition during treatment can be observed in the ratio of accuracy per attempt, which improves for both CF and NOCF participants from session 1 to session 2. This suggests that also the NOCF participants notice errors of V2 and improve their performance during treatment. Though noticing is not directly observable in our setup, learners seemed to have noticed the V2 structure since they applied it correctly more often as treatment progressed. As a result, we can detect learning during treatment as decrease of ApQ, and as the increase of accuracy per attempt. This is relevant for the development of adaptive CALL systems to be able to automatically detect the (proficiency) level of the language learner and the learner's progress.

The findings that proficiency gain seems to be related to different types of learner behaviour in the two conditions may inform the discussion on the role of uptake in language learning. The current findings suggest that learner behaviour can indicate effective practice, but future research could examine in more detail when CF is effective and results in subsequent uptake, to further inform the discussion on the role of uptake for L2 learning (cf. Lyster & Ranta 2013; Goo & Mackey 2013). For example, Loewen (2004) presents convincing evidence of uptake during focus-on-form episodes to be related to learning. In this study we find evidence that uptake supports L2 learning, since the trimmed CF group outperformed the NOCF group.
However, uptake is not a necessity for learning, since some learners in the NOCF group improved without uptake.

5.4. General discussion and future perspectives

The results of the study that we have presented and discussed in the previous sections provide an articulated picture of CF effectiveness and its relationship to learner variables and learner behaviour. It is clear that important questions about CF effectiveness cannot be simply answered with yes or no, but require articulated answers. We see, for instance, that the role of CF and its contribution to language learning may vary depending on the proficiency level of the learner on the specific, targeted grammatical structure. In addition, since the level of proficiency will probably improve as a result of practice, the role of CF is likely to vary accordingly. However, we should note that the effects of proficiency level may not generalize to other learners, e.g. this may be different for child L2 learners, or illiterate or low(er) educated learners.

Insights into how proficiency level interacts with CF effectiveness are important to increase our understanding of language learning, but may pose some problems in terms of application to language pedagogy and classroom instruction, since they require detailed and sophisticated level of analysis of learner behaviour for adaptation to the individual learner. This is hardly feasible in a classroom context. In this connection, the innovative approach adopted in the present study may offer viable solutions. While individualization and adaptation to a single learner may be problematic in a classroom environment, this is clearly less so in a CALL environment. In addition, for speaking practice, Sheen (2008) has shown that learner anxiety can inhibit practice efficiency. In a CALL system, anxiety levels are lower than in face-to-face interaction (Warschauer 1996), suggesting benefits for L2 spoken practice in CALL.

Our research has shown that an ASR-based CALL system can provide personal guidance and assistance to learners by focusing their attention on the targeted structure, by signalling errors, by providing hints, and by having them retry until they produce a correct utterance. Individualisation of instruction has been suggested as an important step forward in L2 learning (Cronbach & Snow 1977; Doughty & Long 2003), and adaptation of CALL systems to learners as an attractive and feasible future direction for CALL (Chapelle 2007; Heift & Rimrott 2012). For instance, a system could automatically detect the most effective type of CF for a particular learner. However, this requires dependable measures of effectiveness, on which the system can base its decisions. So far, the adaptation of systems to learners has been hardly addressed by research (Chapelle 2007). This requires knowledge of specifically which behaviour is indicative of effective practice (i.e. that improves proficiency). Most studies so far have inspected only learner behaviour, or learner performance on the task – while it is necessary to see how learner behaviour relates to learning gain as measured in pre-and post-proficiency tests. This would allow
automatic decisions by CALL systems on the effectiveness of a learner’s practice based on their behaviour, and allow it to adapt to the learner accordingly.

In the present study we have seen how this can be achieved. We have also seen that measures based on modified output (defined as decrease in ApQ over time) could be used in future applications to adapt exercises to the learner’s proficiency level without the need to resort to proficiency tests to determine their level. This could be an easy and simple way to make systems adapt to learners, and the current experiment provides evidence that learner behaviour and proficiency (gain) are related.

The research reported on in this paper indicates future perspectives not only in terms of language didactics and CALL applications, but also with respect to future research on language learning. An ASR-based CALL system as used in this study could be employed to investigate other aspects of feedback and language learning. For instance, it would be interesting to study the effect of CF further by adapting the difficulty level of the questions to the learner’s performance, thereby normalizing for proficiency level. But of course it is also possible to investigate the effect of other forms of CF on spoken language production and learning of other linguistic features.

The current system is developed for language learning and has been shown to be successful in that the learners positively evaluate practice, and improve their proficiency. This is despite that, for experimental rigour, its functionality is highly controlled and options for the learners are restricted. This suggests that these types of systems are well suited for language learning and language learning experiments. Ultimately, it will be interesting to develop an adaptive CALL system that can be implemented in regular language teaching practices, thereby providing ecologically valid learning results, which in turn can be used to further develop CALL systems.

6. Acknowledgements

We would like to thank two anonymous reviewers for their useful comments.
The Role of Automatic CF in Learning L2 Grammar
Chapter 5:
L2 Speaking Practice in CALL: Automatic Feedback and Education Level

This Chapter has been reformatted and slightly modified from:
Bart Penning de Vries, Catia Cucchiarini, Helmer Strik, and Roeland van Hout. Spoken L2 Practice in a CALL environment: the effect of automatic corrective feedback and education level (Submitted to Language Learning and Technology)

Abstract

Studies on second language (L2) learning are conducted mainly with high educated learners, whereas there are reasons to assume that the education background of a learner interacts with L2 learning efficiency. One controversial point in L2 learning is the role of corrective feedback (CF). In order to study the role of CF in relation to individual differences, a controlled experimental environment is necessary. We developed a computer assisted language learning system that can provide consistent CF on spoken production through automatic speech processing. With this system we conducted a language learning experiment, in which we compared 68 learners with different education backgrounds, grouped into Low, Mid, and High educated. The learners were at a similar proficiency level. They were randomly assigned to the CF condition, where they received automatic CF on the grammaticality of their spoken utterances, or to the NOCF condition, where learners practiced speaking with only the option to retry. We find that Mid- and High educated learners benefit from practice, both from the CF condition and the NOCF condition. In line with previous research, we find a trend that the CF condition outperforms the NOCF condition. The low educated learners, however, do not benefit from practice in either condition. Learners all positively evaluated practice with the system. We find differences in practice behaviour between the conditions and the education levels, and relate these to learning outcome. We argue that our study shows that research with high educated learners cannot be generalized to account for all L2 learners, and that for CF research and L2 pedagogy it is important that more effort is invested in recruiting learners of various education backgrounds.
1. Introduction

The role of corrective feedback (CF) in second language (L2) learning is a fundamental topic of discussion in L2 research. The study of CF allows for empirical testing of basic claims in second language acquisition (SLA) theories (e.g. competing claims of universal grammar, cognitive interactionist theory, skill acquisition theory) and has practical and pedagogical applications (R. Ellis 2010).

Corrective feedback is any indication to learners that their utterance contained an error. Because of previously acquired language knowledge (e.g. the L1), particular forms in the L2 may go unnoticed by learners as they lack perceptual salience (Long & Robinson 1998; DeKeyser 2005; N. Ellis 2006), preventing these forms from being acquired. CF can be beneficial to draw attention to these forms.

In meta-analyses of studies on CF (e.g. Russell & Spada 2006; Mackey & Goo 2007; Li 2010), the effectiveness of CF has been shown to vary depending on learner internal variables, and learner external variables. Individual differences have been shown to interact with CF (e.g. proficiency (Mackey & Philp 1998, Anmar & Spada 2006); motivation (Uzum 2010); anxiety (Sheen 2008)). CF effectiveness also varies depending on external factors such as research design (classroom, lab, CALL, face to face), method (outcome measures, CF type, linguistic target), research context (foreign, immersion), and their interaction with individual variables (see Russell & Spada 2006; R. Ellis 2010; Lyster, Saito & Sato 2013 for an overview).

Research on individual differences in CF requires controlling, or accounting for, the learner external variables. As a result there is a call for research settings where the experiment environment can be controlled and clearly operationalized (Li 2010; Russell & Spada 2006; Hulstjin 1997; Goo & Mackey 2013; De Graaff & Housen 2009), and which allow for analysis of how learners interact with CF, to investigate the process of language learning (Mackey & Gass 2006). Thus, besides addressing learning outcomes of CF, studies should also focus on determinants of CF effectiveness (R. Ellis 2009) and see how individual behaviour relates to learning outcomes (Chapelle 2009a).

An important individual difference is education background. Most research so far has been conducted with learners who have a high level of education (Young-Scholten 2013). Reasons for this bias seem to be availability and the additional effort required to design language learning experiments for low educated learners. The ensuing problem is that data is gathered mostly within a specific sub-sample of L2 learners, while the findings are generalized to apply to all L2 learners. In one of the few studies on the effects of education background on L2 learning, the development of the L2 in naturalistic setting was found to depend on education level with lower educated learners remaining lower proficient (Becker et al. 1977, quoted in Young-Scholten 2013). In L2 acquisition (as in L1 acquisition), learners follow a common path in their interlanguage development regardless of education level (Long 1990). However, learners with lower education level appear to develop less far down that common path, and education level may be a factor in explaining that variability (Hawkins 2001). One of the reasons why education level may be a factor could be
that learners with lower education are less able to notice, or benefit from CF. As Tarone and Bigelow (2012) explain: “Making use of corrective feedback requires that the learner engage explicit knowledge about words, their boundaries, and their orders. The literate, educated learners studied in mainstream SLA research have the training to benefit from this kind of corrective feedback” (p.12).

For L2 learning, conscious processes such as noticing are important (Schmidt 1995). Cognitive performance is related to education level (Parisi et al. 2012), and is expected to influence L2 learning and the processing of CF. The noticing of forms in the input or in CF is related to the learners’ processing capacity (R. Ellis 2008). Education level also seems to relate to working memory capacity (Juffs 2006). Working memory is related to language processing and learning (O’Grady 2008) and noticing (Robinson 1995), and it has been shown to influence the noticing of recasts (Mackey, Philp, Egi, Fuji, Tatsumi 2002), also in a CALL experiment (Sagarra & Abbuhl, 2013). In addition, education develops metacognitive skills (Zhang 2010), and research shows that metacognitive skills can influence the learners’ reception of (non-linguistic) CF (Shute 2008). Experimental evidence of the influence of education level and CF comes from Bigelow, Delmas, Hansen, Tarone (2006), where recasts turned out to be ineffective for learners with low literacy and education level. Their explanation was that learners often did not notice the corrective intent of the recasts. Bigelow et al. (2006) replicated the study by Philp (2003) that looked at the interaction of developmental level and noticing of recasts, but with lower educated learners. Compared to Philp (2003), the learners in Bigelow et al. (2006) scored lower on the noticing of recasts.

In response to the call for more controlled studies on CF, we developed a CALL system (‘GREET’) that makes use of automatic speech recognition (ASR) to process spoken language and provide automatic CF. In this way, we could run experiments that kept learner external variables as controlled as possible, while the learners practiced individually with a system that provided consistent CF and recorded learner behaviour. The system was found to be effective, in the sense that learners appreciated practice with the system, it recorded learner behaviour that informed the proficiency analyses, and learners improved their proficiency as a result of practice (Penning de Vries, Cucchiarini, Bodnar, Strik & van Hout 2014). In an experiment comparing a condition that received automatic CF on spoken output with a condition that practiced speaking, we found that learners in the CF condition benefitted most from practice (Penning de Vries, Cucchiarini, Bodnar, Strik & van Hout (forthcoming)). However, this experiment was conducted with high educated learners (N=31, all more than 10 years of formal education, and most attending or having attended university). In light of the evidence that low educated learners with lower education are less able to notice, or benefit from CF. As Tarone and Bigelow (2012) explain: “Making use of corrective feedback requires that the learner engage explicit knowledge about words, their boundaries, and their orders. The literate, educated learners studied in mainstream SLA research have the training to benefit from this kind of corrective feedback” (p.12).

For L2 learning, conscious processes such as noticing are important (Schmidt 1995). Cognitive performance is related to education level (Parisi et al. 2012), and is expected to influence L2 learning and the processing of CF. The noticing of forms in the input or in CF is related to the learners’ processing capacity (R. Ellis 2008). Education level also seems to relate to working memory capacity (Juffs 2006). Working memory is related to language processing and learning (O’Grady 2008) and noticing (Robinson 1995), and it has been shown to influence the noticing of recasts (Mackey, Philp, Egi, Fuji, Tatsumi 2002), also in a CALL experiment (Sagarra & Abbuhl, 2013). In addition, education develops metacognitive skills (Zhang 2010), and research shows that metacognitive skills can influence the learners’ reception of (non-linguistic) CF (Shute 2008). Experimental evidence of the influence of education level and CF comes from Bigelow, Delmas, Hansen, Tarone (2006), where recasts turned out to be ineffective for learners with low literacy and education level. Their explanation was that learners often did not notice the corrective intent of the recasts. Bigelow et al. (2006) replicated the study by Philp (2003) that looked at the interaction of developmental level and noticing of recasts, but with lower educated learners. Compared to Philp (2003), the learners in Bigelow et al. (2006) scored lower on the noticing of recasts.

In response to the call for more controlled studies on CF, we developed a CALL system (‘GREET’) that makes use of automatic speech recognition (ASR) to process spoken language and provide automatic CF. In this way, we could run experiments that kept learner external variables as controlled as possible, while the learners practiced individually with a system that provided consistent CF and recorded learner behaviour. The system was found to be effective, in the sense that learners appreciated practice with the system, it recorded learner behaviour that informed the proficiency analyses, and learners improved their proficiency as a result of practice (Penning de Vries, Cucchiarini, Bodnar, Strik & van Hout 2014). In an experiment comparing a condition that received automatic CF on spoken output with a condition that practiced speaking, we found that learners in the CF condition benefitted most from practice (Penning de Vries, Cucchiarini, Bodnar, Strik & van Hout (forthcoming)). However, this experiment was conducted with high educated learners (N=31, all more than 10 years of formal education, and most attending or having attended university). In light of the evidence that low educated

---

13 Educational attainment is usually associated with greater participation in lifestyle activities that are cognitively demanding. Greater participation in intellectually demanding activities may be especially beneficial for cognitive functions (Parisi et al. 2012).

14 Metacognition refers to knowledge of how to learn, and is found to be an important skill in learning and can compensate for learners with lower aptitude in problem solving tasks (Swanson 1990).
learners process CF differently, we may ask whether the results generalize to a low educated population.

The choice for a CALL experiment environment has consequences for the way in which CF will be perceived (Hulstijn 1997; Lyster et al. 2013). For instance, CF that is implicit in a classroom context is likely to be explicit in a focused learning setting as provided by a CALL system (Heift 2004). However, since CALL is becoming increasingly integrated in L2 teaching practice, it is particularly important to investigate CF in such a context (Chapelle 2009b).

To summarize, there is a call for controlled experiments on CF effectiveness that address the learning process and the learning outcomes, and which take into consideration individual characteristics. One of the individual variables that has not received sufficient attention in the research is level of education. The aim of the present study is to provide an answer to this call. Based on evidence that cognitive performance and metacognitive knowledge differ with education background, we hypothesize that learners with different education backgrounds benefit differently from spoken practice with CF and without CF. To test our hypothesis, we compare learners of different education levels in a controlled experimental L2 learning environment. Through analysis of proficiency, learner behaviour, and learner appreciation we investigate the effect of CF on as a function of education background.

This paper is organized as follows. In Section 2 we describe the method and the ASR-based CALL system we used. Section 3 presents the results obtained for learners of different education levels. In section 4 we discuss the results and draw conclusions.

2. Method

2.1. Participants

We recruited 68 language learners in four different locations from classes teaching A1-A2 CEFR level Dutch: the high educated learners (High-Ed) were mostly from the university language centre, while the lower educated learners (Mid- and Low-Ed) came from the three other locations, where we had a high drop-out rate. This was because the sessions were (necessarily) scheduled less flexibly, and learners more often did not complete the exercise.15 On average, the learners were 32 years

---

15 At the three Low- and Mid-Ed locations, 59 learners initially (were) signed up, of whom 28 completed pre- and post-tests. The dropout rate can be explained by several factors: (1) the experiment was part of the course, but attendance was not obligatory; (2) the time schedule was inflexible (at the university, if someone was unable to attend, a different time could be arranged); (3) the experiments were in fixed time slots, but Mid- and Low-Ed learners generally went through practice slower. Especially when learners came in late, they would have insufficient time to complete all exercises. Only completed practice sessions were included in this study.
old (SD=8.7; min 18 max 62). They received 15 euro for completing two experiment sessions and were randomly assigned to an experimental condition: either the CF condition, where learners receive automatic CF on their grammar in a spoken utterance (N=40), or the NOCF condition, where learners practice speaking with the option to self-correct, but without receiving CF from the system (N= 28). The skewed distribution is caused by the inclusion of nine High-Ed CF participants from an experiment with two types of CF: recasts and prompts. As the CF (prompt) condition from that experiment is identical to the CF group of the other experiments, we included these learners.

2.1.1. Education level

On the pre-test questionnaire, the participants indicated their level of education. They selected from a list of six options, where one category (‘other’) allowed for specification of an education level that was not on the list. These were assigned to the appropriate category by the experimenters. We assigned the education levels in the present study as follows:

- Low-Ed: completed at most primary school (N=12)
- Mid-Ed: completed at most secondary school (N=15)
- High-Ed: completed over six years of secondary school (N=41)

2.1.2. Language knowledge

The 68 participants have 31 different L1s. Per education subgroup, there are 10, 12, and 24 L1 types for Low-, Mid-, and High-Ed, respectively. Knowledge of other languages next to the L1 may also influence Dutch learning. On average, 74% (50/68) of the participants spoke an L2 next to Dutch: for Low-Ed 58% (7/12), Mid-Ed 40% (6/15), and High-Ed 90% (37/41). Table 1 gives the languages spoken by the learners.

Some of the L1 and L2 languages of the participants have verb second (V2), which is the target structure in the current experiment (see Section 2.4.1): Danish, German, and Swedish. Transfer from L1 is expected, and transfer is possible from the L2 depending on whether the structure has been learned in the L2 (i.e. dependent on proficiency level) (Odlin 1989; Bohnacker 2006; Ringbom 1992). The issue of transfer is addressed by taking pre-test proficiency into account (Section 3.2).

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2 (next to Dutch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Ed</td>
<td>Arabic (2), Berber, Chinese, English,</td>
<td>Arabic, English (5), Galician,</td>
</tr>
<tr>
<td>(N=12)</td>
<td>Farsi, Greek, Polish, Pashto, Spanish,</td>
<td>Portuguese, Spanish, Swahili (2),</td>
</tr>
<tr>
<td></td>
<td>Thai (2)</td>
<td>Swedish, none (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mid-Ed</td>
<td>Arabic (3), Dari, English, Farsi (2),</td>
<td>English (4), French, Spanish, none (9),</td>
</tr>
<tr>
<td>(N=15)</td>
<td>French (2), Italian, Indonesian, Kurdish,</td>
<td>unspecified (1)</td>
</tr>
<tr>
<td></td>
<td>Polish, Thai, Turkish</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. Language background of the participants in the respective education levels. If more than one, the number of speakers are given in brackets.

2.2. Procedure

The experiment took place on two separate days within a week. When participants came in, they would first log in to the experiment website, which presented them sequentially with their tasks for the day. On day 1, the participants filled in a short questionnaire about their background, and completed a discourse completion task (DCT) to measure pre-test proficiency (see Section 2.5). The DCT was preceded by click-through instruction web pages and three trial items. The test took 10-15 minutes to complete. After the DCT, the participant moved on to 45 minutes of treatment (Section 2.4), which was also preceded by click-through instruction pages. On day 2, the participants started with treatment (45 minutes), completed the DCT, and finally filled out a post-test questionnaire (Section 2.6). In total, each session took 75-100 minutes. All interactions of the learner with the system were recorded, in order to inspect learner behaviour during treatment (Section 2.7).

The system and procedure were piloted with 12 High-Ed learners and after a first experiment demonstrated the system to be effective with High-Ed learners (Penning de Vries et al. 2014), we piloted the system with 11 Low-Ed participants. Minor alterations (to pre-test tutorials, added a microphone check, and to user interface) were made as a result. All learners in the current experiment worked with that system.

2.3. The FASOP (GREET) system

This study uses a CALL system, GREET, that provides CF on spoken output through ASR. Learners work individually and complete grammar exercises in the form of interactive dialogues. This method allows for a detailed and individual approach that is highly consistent across multiple participants. For a detailed description of the system we refer to (Penning de Vries et al. 2014, and for the ASR to Van Doremalen, Cucchiarini & Strik 2010).

The GREET program is accessed online through a web browser. The participants log in and interact with the system through a graphical user interface. They record spoken utterances by combining words graphically presented in ‘word
blocks’. The speech recognizer maps the recorded utterance to a possible order of the word blocks (the order that is most probable). All orders of the word blocks are tagged for grammaticality, which the system uses to provide feedback. All speech recognition is performed on a central web server. This allows us to run experiments simultaneously, at different locations, and to store all data centrally. In addition, the GREET system logs system-learner interactions for analyses of learner behaviour and system performance.

2.4. Treatment

Participants watched short clips (30–45 s) of a video developed by an educational publisher about a man moving into a new apartment and meeting his neighbours (Put & Peekel 2001). After each video they were asked three to five questions about the clip. Responses were to be formed by using word blocks (segments of the sentence) presented on the screen in random order. The learner would then press record and speak the sentence (note that the learner cannot drag-and-drop the blocks).

The treatment is designed in two parts. The first session contains 65 questions (25 target, 40 distracters); the second session contains 59 questions (19 targets, 40 distracters). Learners vary in their speed of practice, and some complete all questions of a session within the 45 minutes of practice. To control for time-on-task, learners automatically restarted to redo questions to complete 45 minutes of practice.

2.4.1. Target structure

The target structure is inversion of subject and verb in declarative main clauses, as a result of verb second (V2) properties of Dutch. V2 entails that the finite verb appears in second position regardless of the first constituent in the main clause. The subject precedes the finite verb when it is the first constituent. If the first constituent is not the subject, the subject follows the finite verb immediately. This is illustrated in Table 2.

<table>
<thead>
<tr>
<th>Target structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Subject-initial main clause</td>
</tr>
<tr>
<td>Jan  \hspace{0.5cm} Koopt  \hspace{0.5cm} Morgen  \hspace{0.5cm} bloemen</td>
</tr>
<tr>
<td>Jan \hspace{0.5cm} buys (3SG) \hspace{0.5cm} tomorrow \hspace{0.5cm} flowers</td>
</tr>
<tr>
<td>Subject \hspace{0.5cm} Verb \hspace{0.5cm} Adverbial \hspace{0.5cm} Object</td>
</tr>
<tr>
<td>(2) Adverbial-initial sentence: Postverbal subject</td>
</tr>
<tr>
<td>Morgen \hspace{0.5cm} Koopt \hspace{0.5cm} Jan \hspace{0.5cm} bloemen</td>
</tr>
<tr>
<td>Tomorrow \hspace{0.5cm} buys (3SG) \hspace{0.5cm} Jan \hspace{0.5cm} Flowers</td>
</tr>
<tr>
<td>Adverbial \hspace{0.5cm} Verb \hspace{0.5cm} Subject \hspace{0.5cm} Object</td>
</tr>
</tbody>
</table>

*Table 2. An illustration of the target structure. The subject-verb order changes to comply with verb second when the first constituent of the sentence is not the subject.*
The acquisition of V2 is problematic for L2 learners of Dutch (Jordens 1988). Violation of V2 does not affect the meaning of the sentence, lacks a clear form-meaning relationship, and is unlikely to lead to communication breakdowns. The error often remains in the interlanguage for a long time. This makes the feature an appropriate candidate for studying CF, as it may not be noticed in meaning-based communication, and a focus on language form is necessary.

2.4.2. Corrective feedback in our CALL system

In the CF condition, prompt feedback explicitly told learners whether their utterance was correct or incorrect. For a correct answer, the learner was shown a green check mark and moved to the next question. For an incorrect answer, the learner was shown a red screen saying ‘That is incorrect. Please try again’ and returned to the same question. After the second error, the system gave a hint in the form of a word block placed in its appropriate place in the answer sentence. To prevent frustration (due to either learner error, or ASR error) the system made a skip-button available after four wrong tries. In the NOCF group, after each recording by the learner, the system gave them the following message in a white pop-up screen: “You can only save one answer. Do you wish to save this one and move to the next question?” They then had the options: (1) retry, or (2) move to the next question. They were promised a score at the end of practice.

2.5. Proficiency test: discourse completion task

We employed a discourse completion task (DCT) to measure pre-post-test changes in spoken proficiency. The participants were asked to complete an unfinished sentence on the screen, and record the spoken sentence. For each item, they were given a lead-in of the sentence, and a noun and/or an unconjugated verb. A picture was shown for context. The test was counterbalanced in two versions, and had 16 target items and 16 distracter items. Distracter items contain common errors by L2 learners of Dutch. The order of items is randomized per participant. For answering each item there is a time limit of 30 seconds. If the participant does not record an answer within that time, the response is scored as incorrect.

2.6. Learner evaluation questionnaire

At the end of the second session participants completed a questionnaire to evaluate the system and their learning experience (see Appendix H). There were 19 CF related and 17 NOCF related questions that were answered on a semantic differential 5-point scale. Nine questions about usability, enjoyment, and perceived effectiveness of the program were shared for CF and NOCF. The other questions concerned preferences for modality and CF, and control questions. In addition, two
open questions asked about preferences for feedback, and there was room for
general comments.

2.7. Logged data of learner behaviour

Through the logs of learner-system interactions we have access to information about
time, and duration (e.g. log-in time, and time spent on each page), and about the
specific interactions related to the practice exercise, (e.g. the number of attempts per
question). System actions are also logged (e.g. the recognition result of the learner’s
utterance, and the CF presented). Most relevant for this paper are the number of
questions practiced, the number of attempts at those questions, and the number of
times learners produced a correct utterance.

3. Results

3.1. Participant background

We applied a one-way ANOVA (between subjects factor education level) to the
following items in the pre-questionnaire: (a) length of residence in the Netherlands
(b) time learning Dutch, (c) time using the computer, (d) time reading/writing
Dutch, and (e) time listening/speaking Dutch. No significant difference was found
for items (d) and (e): the time per day that the learners were exposed to, or used
Dutch, was similar across education levels. Means for (a,b,c) are shown in Table 3.

<table>
<thead>
<tr>
<th>Education level</th>
<th>Length of residence (years) (mean (SD); min-max)</th>
<th>Time learning Dutch (years) (mean (SD); min-max)</th>
<th>Time using computer (Days per week) (mean (SD); min-max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Ed (N=12); 4f, 8m</td>
<td>4.5 (4.47); 1 – 13</td>
<td>2.4 (2.20); .42 – 8</td>
<td>4.8 (2.58); 0 – 7</td>
</tr>
<tr>
<td>Mid-Ed (N=15); 7f, 8m</td>
<td>6.1 (8.60); .25 – 25</td>
<td>1.0 (.78); .13 – 3</td>
<td>5.1 (2.54); 1 – 7</td>
</tr>
<tr>
<td>High-Ed (N=41); 25f, 16m</td>
<td>1.1 (1.14); .17 – 4</td>
<td>0.6 (.76); .1 – 4</td>
<td>6.5 (1.25); 3.5 – 7</td>
</tr>
<tr>
<td>Total</td>
<td>2.8 (4.82); .17 – 25</td>
<td>1.0 (1.33); .1 – 8</td>
<td>5.9 (1.99); 0 – 7</td>
</tr>
</tbody>
</table>

Table 3. Learner background characteristics per education level, based on
information provided by the participants in a pre-questionnaire.

There is a significant difference in the length of residence (F(2,66)=7.795, p=.001).
The Mid-Ed have lived the longest in the Netherlands, and the High-Ed the shortest
The Role of Automatic CF in Learning L2 Grammar

(post hoc test, Sidak); the difference between the Low-Ed and High-Ed was not significant. The reasons for coming to the Netherlands complement the length of residence data: In all three education groups, approximately half of the learners have come to the Netherlands for a/their partner. For High-Ed learners, ‘study’ was a common reason (29%), but not for the Mid-Ed (7%) or Low-Ed (8%). Among the Low-Ed, ‘work’ was a common reason (42%) versus 13% for Mid-Ed and 20% for the High-Ed. Four participants had come as refugees (1 Mid-, and 3 High-Ed), and five participants had not specified a reason (1 Low-, 3 Mid-, 1 High-Ed).

The time spent learning Dutch is different between education levels (F(2,66)=12.165, p=.000). The High-Ed learners have spent significantly less time learning Dutch than the Mid- and the Low-Ed learners (post hoc test, Sidak).

Finally, there is a significant difference between the education levels on computer use (F(2,67)=3.468, p=.005). The High-Ed participants use the computer more often than the Mid- and the Low-Ed (Post hoc test, Sidak).

3.2. Discourse completion task (DCT)

The DCT has a high reliability score (32 items, Cronbach’s alpha = .954). Figure 1 shows the individual scores, labelled per education level, and split out for the CF and NOCF condition. The participants’ pre-test scores are seen to range from zero to full proficiency. The respective CEFR levels A1 and A2 correspond to means of DCT scores (A1: M=.26 (SD=.293); A2: M=.30 (SD=.316)), but we find at the same time that there is a wide range of scores both for A1 (.00 -.94) and A2 (.00-1.00) learners. In other words, there is much variability of V2 proficiency within learners of the same proficiency level. Within each education level we find comparable V2 proficiency ranges: Low-Ed .00-.81; Mid-Ed .00-.94; High-Ed .00-1.00. As a result, high and low V2 proficient learners are found in all education background groups.

To control for ceiling effects, we excluded learners with a high pre-test score. We set the limit of inclusion for analysis at DCT pre-test score <=.60. There are three main reasons for this: 1) ceiling level learners cannot learn from CF nor from NOCF, and their behaviour during treatment will not inform us on learning behaviour; 2) in our sample there are participants with mastery of a V2-language. For participants with an L1 containing V2, it is justified to exclude them categorically, as they will all know V2. For participants with an L2 with V2, exclusion should be based on whether they have knowledge of V2 in that L2. We cannot categorically exclude these learners, but should exclude them on the basis of their DCT pre-test score; 3) generally learners with proficiency levels over DCT pre-test=.60 were found to complete all the questions in the treatment well within 90 minutes, which suggests that the treatment was easy for them. The limit causes 15 participants to be excluded:

- 2 Low-Ed (one lived for 13 years in NL; one makes a large unexplained decrease in score)
- 3 Mid-Ed (one lived for 25 years in NL)
• 10 High-Ed (three German L1, four German L2, one Swedish L2, one with a large unexplained decrease in score)

Figure 1. Scatter plot (jittered) showing the individual DCT scores. The participants are split out by condition (CF/NOCF) and education level. The x-axis gives the pre-test score, the y-axis gives the post-test score. The dashed line indicates the cut-off limit, the solid diagonal line is for reference.

The mean scores of the remaining 53 participants (Low-Ed: 5 CF, 5 NOCF; Mid-Ed: 5 CF 7 NOCF; High-Ed 21 CF, 10 NOCF) are given in Figure 2, split out by education level.
Figure 2. Line graph of the mean proportion correct on the DCT pre- and post-test. The participants are split out by condition and education level.

Figure 2 suggests differences in learning trajectories. To test for statistical significance we performed an ANOVA with two between subjects factors: education level (Low, Mid and High) and condition (CF and NOCF) and one within subjects factor (time, pre-test versus post-test). The results are given in Table 4.

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>F(1,47)=8.380, p=.006</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of time</td>
<td>F(1,47)=1.756, p=.192</td>
</tr>
<tr>
<td>Interaction time by condition</td>
<td>F(2,47)=3.293, p=.046</td>
</tr>
<tr>
<td>Interaction time by education level</td>
<td>F(2,47)=.015, p=.985</td>
</tr>
<tr>
<td>Interaction time by condition by education level</td>
<td>F(2,47)=.219, p=.804</td>
</tr>
</tbody>
</table>

Table 4. Results of an ANOVA on DCT results, with factors time, condition, and education level.

The results show that there is a significant effect of time, education, and interaction of time by education. There is an overall improvement, but there is a difference in improvement according to education level. As seen in Figure 2, the Low-Ed learners do not improve, whereas the Mid- and High-Ed seem to improve. Because there is an interaction effect of time by education level, we applied ANOVA on the learners split by education level. The results are given in Table 5.
Table 5. Results of a repeated measures ANOVA of the DCT scores, split by education level.

The results in Table 4 confirm that the Low-Ed participants do not improve as a result of treatment, irrespective of the CF or NOCF condition. Both the Mid-Ed and the High-Ed group improve as a result of treatment. However, within the education levels there is no significant interaction of condition with time. In other words, we find no difference in the amount of improvement between the CF and the NOCF condition. In an earlier study we found a positive effect of CF, but this sample consisted of High-Ed learners (27 High-Ed, 4 Mid-Ed). In a combined analysis of Mid- and High-Ed of the current sample, we find a trend in accordance with that finding (interaction effect of time by condition: F(1,41)=3.274, p=.078, \( \eta^2=.074 \)).

3.3. Learner evaluation

The post-test questionnaire items are reliable (9 items, Cronbach’s alpha = .774), and we group them to give a mean score. Inspection of individual items did not reveal any particularly low-scoring items. The mean scores of the learners’ evaluation are given in Table 6, split out for the three education levels and the two treatment conditions.

Learners evaluated the system positively in both conditions. A two-way ANOVA (between subjects factors: condition and education level) showed no difference between the conditions (F(1,51)=1.437, p=.237), no effect of education level (F(2,51)=.286, p=.752), and no interaction of condition by education level (F(2,51)=.689, p=.507).

On the DCT, we found that the Mid-Ed and the High-Ed improved after treatment, whereas the Low-Ed did not. Interestingly, though the Low-Ed did not improve, they were positive about practice with the system. In a correlation analysis, no relation shows up between the V2 proficiency test results (DCT) and the learners’ ratings of practice with the system.\(^{16}\)

\(^{16}\) In addition, learner ratings did not correlate with learner behaviour data during practice (reported in Section 3.4)
120 | The Role of Automatic CF in Learning L2 Grammar

<table>
<thead>
<tr>
<th></th>
<th>CF M (SD)</th>
<th>NOCF M (SD)</th>
<th>Total M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Ed</td>
<td>3.87 (.499), N=5</td>
<td>3.56 (.691), N=4</td>
<td>3.73 (.573), N=9</td>
</tr>
<tr>
<td>Mid-Ed</td>
<td>3.67 (.673), N=4</td>
<td>3.73 (.867), N=7</td>
<td>3.71 (.770), N=11</td>
</tr>
<tr>
<td>High-Ed</td>
<td>4.04 (.449), N=21</td>
<td>3.62 (.413), N=10</td>
<td>3.90 (.474), N=31</td>
</tr>
<tr>
<td>Total</td>
<td>3.96 (.488), N=30</td>
<td>3.65 (.615), N=21</td>
<td>3.83 (.560), N=51</td>
</tr>
</tbody>
</table>

Table 6. Means and standard deviation of the scores on the post-test questionnaire. Participants rated practice with the system on a 5 point semantic differential scale: 5 is most positive, and 1 is most negative.

3.4. Learner behaviour

The log data provides information about learner behaviour during practice. For four subjects (High-Ed, CF) log data was missing or incomplete. They were excluded in this part of the analysis. We focused on the following data available in the logs:

- Questions: the number of times a V2 sentence was practiced at least once.
- Attempts: the amount of produced output by the learner. Per question, a learner can make multiple attempts. The attempts include errors, reformulations (modified output) and correct productions of a V2 sentence.
- OKs: the number of correct V2 utterances produced.

To determine grammatical accuracy during practice, we calculated two ratios:

- Attempts per question (ApQ). This ratio should be interpreted differently for the CF and the NOCF condition. In the NOCF condition, the number of attempts is self-regulated: learners can proceed to the next question regardless of the correctness of the answer. This ratio therefore informs us mostly on how motivated the learner may have been to achieve a correct score. In the CF condition, the system explicitly signals the error, and prompts the learner to try again (with each subsequent error they receive a hint) until they answer correctly. The ratio thus indicates a) the number of CF messages and hints from the system that the learner received, and b) the number of tries by the learner to produce the correct utterance.
- OK per Attempt: for both the CF and the NOCF condition, the ASR assesses the grammaticality of the learners’ responses, thus we can calculate the ratio of correct utterances produced. In the CF condition, the learners must produce a correct utterance (OK) to proceed, whereas this is optional in the NOCF condition. However, we can compare the change in this ratio during practice.

Figure 3 gives an overview of the log data available, split out for session, level of education, and condition.
Figure 3. Bar graphs with standard errors (± 2 SE) of the number of Questions (Q), Attempts (A) and correct utterances (OK) per session (S1, S2).

We applied a repeated measures ANOVA with two between subjects factors (condition and education level) and one within subjects factor (time, session 1 versus session 2). The results are given in Table 7. The results show a consistent difference between the practice conditions: the learners in the NOCF condition practice more questions than the CF group, but the CF group produces more attempts, and more OKs.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Number of Questions</th>
<th>Number of Attempts</th>
<th>Number of OKs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within subjects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>F(1,39)=3.354, p=.075</td>
<td>F(1,39)=.328, p=.570</td>
<td>F(1,39)=1.018, p=.319</td>
</tr>
<tr>
<td>Time by Condition</td>
<td>F(1,39)=6.289, p=.016</td>
<td>F(1,39)=3.756, p=.060</td>
<td>F(1,39)=.783, p=.382</td>
</tr>
<tr>
<td>Time by Education level</td>
<td>F(2,39)=.871, p=.427</td>
<td>F(2,39)=4.499, p=.017</td>
<td>F(2,39)=.125, p=.883</td>
</tr>
<tr>
<td>Time by Condition by Education level</td>
<td>F(2,39)=2.379, p=.106</td>
<td>F(2,39)=6.763, p=.003</td>
<td>F(2,39)=.514, p=.602</td>
</tr>
<tr>
<td><strong>Between subjects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>F(1,39)=7.458, p=.009</td>
<td>F(1,39)=20.280, p=.000</td>
<td>F(1,39)=32.193, p=.000</td>
</tr>
<tr>
<td>Education level</td>
<td>F(2,39)=1.913, p=.161</td>
<td>F(2,39)=.519, p=.599</td>
<td>F(2,39)=3.837, p=.030</td>
</tr>
<tr>
<td>Condition by Education level</td>
<td>F(2,39)=1.789, p=.181</td>
<td>F(2,39)=1.095, p=.345</td>
<td>F(2,39)=832, p=.443</td>
</tr>
</tbody>
</table>

Table 7. Results of a repeated measures ANOVA of log data learner behaviour.
For the number of questions practiced, there is a significant interaction effect of time by condition. As Figure 3 shows, the number of questions practiced is stable in the CF condition, whereas it decreases in the NOCF condition. This is probably due to (a) the number of target questions: 25 in session 1, and 19 in session 2, and the fact that (b) the learners in the NOCF condition are not slowed by feedback, often complete all questions and have to restart. In the second session they are more aware of the length of the content and pace themselves.

The CF condition gives more attempts (effect of condition). As the practice time is equal for all participants this means that the learners in the CF condition produce attempts at a higher rate. The number of attempts changes differently per education level and condition (effect of interactions). This can be seen in Figure 3: in the CF condition, the Low-Ed seem to make more attempts, whereas the Mid-Ed are stable, and the High-Ed decrease. Low-Ed learners seem to require more time to get familiar with practice; they also asked for more assistance to make sure they were doing it correctly (as will be discussed below, Low-Ed learners made more attempts per question in the second session, suggesting that they were less anxious about making mistakes). High-Ed learners seem to be more used to working independently with a computer program, which suggests that the increase has to do with practice familiarity. The decrease for the High-Ed is related to content length: there are fewer questions in the second session.

Learners in the CF condition produce more OKs than learners in the NOCF condition. In addition, there is an effect of education level. A post hoc test (Sidak) shows that the difference is significant between Low and High, and Mid and High. As seen in Figure 3, the High-Ed learners produce more OKs than the Low- and Mid-Ed learners.

Figure 4. Bar graph with standard errors (± 2 SE) of the ratios of Attempts per question (ApQ) and OKs per Attempt (OKpA) in session 1 (S1) and session 2 (S2). The participants are split out by condition and by education level.
Learner behaviour is found to differ per condition, for number of questions practiced, number of attempts and OKs. Education background has an effect on the OKs. We turn to the ratios of attempts per question and the OKs per attempt to look in more detail at changes in performance over time. These are given in Figure 4.

The results of a repeated measures ANOVA with condition and education level as between-subjects factors and time as within-subjects factor (time, session 1 versus session 2), are given in Table 8.

For the attempts per question there is a clear difference between the two conditions: learners in the NOCF condition generally do not retry, whereas the CF group practices each question more than twice, as a result of CF forcing them to retry after an error. The function of the CF should be to draw attention to the error, and to help the learner repair the error. If the CF succeeds in this goal, the number of ApQ over time should decrease. For the group overall, the ANOVA results show a non-significant change in APQ over time. However, there is a significant effect of the three-way interaction of time by condition by education level, and a significant effect of the interaction of condition by education level. Figure 4 suggests that in the CF condition, the High-Ed learners decrease in ApQ, Mid-Ed learners are stable, and Low-Ed learners increase. To test this assumption, we performed an ANOVA on the participants split out by education level. We found a non-significant effect of time (session 1, session 2) for the Low- and Mid-Ed learners (for both F<1), and a near significant effect of time (F(1,25)=3.225, p=.085) for the High-Ed learners.

<table>
<thead>
<tr>
<th>Effect</th>
<th>ApQ</th>
<th>OKs per Attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within subjects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>F(1,39)=.040, p=.842</td>
<td>F(1,39)=2.599, p=.115</td>
</tr>
<tr>
<td>Time by Condition</td>
<td>F(1,39)=.057, p=.813</td>
<td>F(1,39)=0.086, p=.771</td>
</tr>
<tr>
<td>Time by Education level</td>
<td>F(2,39)=3.092, p=.057</td>
<td>F(2,39)=2.115, p=.135</td>
</tr>
<tr>
<td>Time by Condition by Education level</td>
<td>F(2,39)=4.042, p=.025</td>
<td>F(2,39)=.959, p=.392</td>
</tr>
<tr>
<td><strong>Between subjects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>F(1,39)=79.099, p=.000</td>
<td>F(1,39)=19.511, p=.000</td>
</tr>
<tr>
<td>Education level</td>
<td>F(2,39)=.841, p=.439</td>
<td>F(2,39)=6.262, p=.004</td>
</tr>
<tr>
<td>Condition by Education level</td>
<td>F(2,39)=3.999, p=.026</td>
<td>F(2,39)=.213, p=.809</td>
</tr>
</tbody>
</table>

Table 8. Results of a repeated measures ANOVA on the Attempts per Question (ApQ) and the OKs per Attempt with the between subjects factors condition (CF, NOCF) and education level (Low, Mid, High).

In the CF condition, the change in ApQ over time differs per education level. The High-Ed group suggests a decrease in ApQ. This is clearly not the case for Low-Ed learners, who seem to increase, and for the Mid-Ed learners, the ApQ seems stable.
In the NOCF condition, learners generally do not retry (ApQ=1), though Figure 4 suggests that High-Ed learners sometimes retry.

To complement the ApQ, we analysed the OKs per attempt. Since the NOCF participants generally did not reformulate, there are no changes over time. To know if the performance of the NOCF participants improved over time, we can inspect the OKs per attempt as a measure of the accuracy of the responses (cf. Penning de Vries et al. forthcoming).

The results of an ANOVA in Table 8 show no overall effect of time for the OKs per attempt. There is an effect of condition: in the CF condition, learners are alerted to errors, and required to produce an OK before they can move on to the next question. As a result, the OK,pA. is higher overall in the CF condition. In addition, there is a difference between the education levels: a post hoc test (Sidak) shows the Low-Ed and Mid-Ed groups having lower OKs per attempt than the High-Ed.

To inspect the difference in education level in more detail, we ran an additional ANOVA on the participants split out by education level. Here we find that for the Low- and the Mid-Ed learners there are no significant effects of time (both F<1), no significant interaction effects (both F<1), but a significant effect of group (Low: F(1,7)=76.999, p=.000 and Mid: F(1,9)=23.063, p=.001). For the High-Ed learners there is a significant improvement over time (F(1, 24)=14.415, p=.001), with no interaction effect (time by condition F(1,24)=.203, p=.666), and an effect of condition (F(1,24)=9.880, p=.004).

The ratio of OKs per Attempt was inspected to detect changes in learners’ accuracy during practice. We find that learners in the CF condition have a higher ratio of OKs per Attempt than those in the NOCF condition. There is an effect of education level: High-Ed learners have a higher ratio overall than Mid- and Low-Ed. In addition, High-Ed learners are found to improve their accuracy during practice, in the CF and the NOCF condition. This is not found for the Low- and Mid-Ed learners.

4. Discussion and conclusion

Our study compared the language learning results and learner behaviour of L2 learners with various education backgrounds who used the same CALL system that employs ASR technology to provide practice and feedback on speaking performance. We collected data with two different configurations of the system, the CF and NOCF conditions, to investigate whether the effect of CF is related to education level. Learners in the CF condition received automatic CF, while learners in the NOCF condition practiced without feedback, but were promised a score after practice.

The results of the spoken proficiency test show that education level interacts with practice effectiveness. Learners with Mid- and High education backgrounds improved in both the CF and the NOCF condition. The Low-Ed group did not improve, in neither condition.
In an overall comparison of the CF and the NOCF conditions (i.e. including all subjects), there was no difference in learning outcome. In an earlier experiment (Penning de Vries et al. forthcoming), the CF condition outperformed the NOCF condition when excluding high-proficient learners. In this paper, we set the same upper limit of pre-test proficiency. We found the same trend of the CF condition outperforming the NOCF condition among the Mid- and High-Ed learners combined: when excluding Low-Ed learners. This emphasizes that findings of CF experiments with high educated learners cannot straightforwardly be generalized.

Learners in all education groups positively evaluated practice with the system. It was not the case, therefore, that Low-Ed learners did not value working with the system, and that that influenced learning effectiveness. The reverse is also true: the effectiveness of the system in improving the learners' V2 proficiency did not affect the learners' appreciation.

In order to explain the difference in effectiveness of practice, we investigated learner behaviour during practice. In the NOCF condition, learners practiced more questions than in the CF condition, and thus practiced with more instances of V2. Higher educated (Mid- and High-Ed) learners benefitted from the NOCF condition, but lower educated did not. We assume that the effectiveness of the NOCF condition relies for a large part on the learner. The exercise draws some attention to form by asking learners to construct a sentence, and thus they must think about word order. It requires the learner's effort to produce a grammatically correct sentence, and subsequently additional effort to reflect on the utterance and potentially correct a V2 error. In addition, this requires the ability of the learner to produce and notice correct use of V2.

With respect to effort, we find that learners in the NOCF condition produce fewer attempts than learners in the CF condition in the same practice time. This means that learners produced attempts at a slower rate than the CF learners. We assume that this is related to the fact that learners in the NOCF condition think longer before they attempt as they know they will not receive CF, and thus want to record a correct attempt at the first go. This also suggests that they are more cognitively involved in answering, which is beneficial for language practice (cf. R. Ellis 2008). Between education levels, we do not find a difference in the number of attempts, which suggests that all learners spend a similar amount of effort (in terms of output).

There are two differences in learner behaviour that suggest that the learners were differently concerned with the accuracy of their responses (which we tried to stimulate by promising a delayed score). We find that only the High-Ed learners sometimes retried a question. This means that they perceived an error and were concerned with accuracy, or their score. We also find that the Low-Ed learners practiced the highest number of questions (though a non-significant difference). This suggests that they went through the questions more quickly than the Mid- and High-Ed learners, and that they did not invest much effort in trying to answer the questions correctly. Lower educated learners were perhaps less concerned with L2 accuracy, and more with communicative adequacy, and therefore quickly regarded
their answer as correct. On the other hand, the high number of questions practiced may also relate to an overconfidence in their performance. Kruger and Dunning (1999) found that learners with lower metacognitive skills are more likely to overestimate their ability.

With respect to ability, the trend of more effective practice behaviour for the High-Ed learners may be related to higher metacognitive skills. Because there is no clear difference in behaviour between the Mid- and the Low-Ed learners in the NOCF condition, while there is a difference in learning outcome, it seems likely that cognitive performance influenced the effectiveness of practice. Lower educated learners (on account of lower cognitive skills) are less likely to notice and attend to errors (cf. Mackey et al. 2002). By inference, they may be less skilled in detecting the rule underlying the error.

Learners appear to prefer CF in the classroom (Lyster et al. 2013). Kartchava (2012) writes that “the students whose instructional expectations are not met may consciously or subconsciously question the credibility of the teacher and/or the instructional approach in cases where corrective feedback is not provided […], which may further lead to a decrease in learner motivation and the amount of learning achieved.” (p.113). In the current study, however, we find that the NOCF condition was positively evaluated and not significantly lower than the CF condition. The task struck the learners as useful practice, also without CF. Our previous studies, with high educated learners, found that the CF condition was rated more positively (Penning de Vries et al. 2014, forthcoming). In our present study we found a trend: the High-Ed learners seemed to prefer the CF condition; the Mid-Ed and Low-Ed showed no preference for either condition. Explanations for why only higher educated learners seem to prefer the CF condition may be: a greater concern for accuracy; some learners in the NOCF condition (Mid- and Low-Ed) thought that practice was for listening skills, which renders CF or NOCF moot; or the preference for (this type of) CF is not as strong among Low-Ed learners. In CALL, closer (situated) monitoring of motivation is possible, and this may be necessary to detect these differences in learner attitude towards CF and NOCF (cf. Bodnar, Cucchiarini, Strik & van Hout 2014).

Learners in the CF condition produced more attempts and OKs than learners in the NOCF condition. So they produced more spoken Dutch and more grammatical instances of V2. In addition, they received CF, and thus also input regarding V2. High- and Mid-Ed learners improved as a result of practice, while the Low-Ed learners did not.

There was no significant difference between the education levels in the number of attempts made, thus the practice effort was comparable. Between the education levels, the High-Ed learners were found to produce most OKs. Following L2 learning theories that draw on skill acquisition (cf. De Bot 1996; DeKeyser 2007), production of grammatical output facilitates the automatization of these forms, allowing for more rapid access over time, and thus improving proficiency. Thus, Low- and Mid-Ed learners had fewer opportunities for proceduralization of V2. However, Mid-Ed learners improve on the post-test, so the reason for non-
improvement in the Low-Ed learners is not solely the number of OKs. There are reasons to assume that lower educated learners need more exposure for similar learning results. This was found for verbal learning by Meijer, De Groot, van Boxtel, van Gerven, Jolles (2008).

Besides the benefit of producing the correct form for proceduralization, an OK also offers opportunities for noticing V2 in the input (i.e. when given by the program as a hint), or in the spoken output (by the learner). This noticing should be facilitated by the CF, which explicitly notifies them of errors (but does not pinpoint it). We found indications that the High-Ed learners pick up the V2 feature during practice: their number of CF instances (as seen in the ApQ) seemed to decrease, and the percentage of correct attempts improved over time. So their behaviour shows uptake of the V2 feature. This was not found for the Mid- or the Low-Ed group. Interestingly, however, Mid-Ed learners did improve their proficiency in the post-test.

In learner behaviour, there are no significant differences between the Mid- and Low-Ed learners. We assume that differences in cognitive and metacognitive resources may explain that Mid-Ed learners benefit more from a similar amount of CF, modified output, repair, and OKs. Although repair of an error may indicate that the feature is noticed (Hegelheimer & Chapelle 2000; Chapelle 2009b), this is not necessarily the case. Repair should be tracked over time, to detect uptake of the target feature.

As the amount of CF is comparable across education levels (as seen in the ApQ), possible explanations should be sought in the nature of the CF provided. It is possible that the type of CF provided in our CALL system is not effective for Low-Ed learners. Although the practice behaviour of the Mid-Ed learners does not indicate learning over time, they did improve on the post-test. Research indicates that learners need to notice which error is targeted for CF to be beneficial (El Tatawy 2002). A possible explanation is then that the Low-Ed learners apparently need more guidance to notice the V2 feature. For instance through more explicit CF that draws attention to the error and provides the correct form or the rule; or by alerting learners before the experiment to the V2 feature (this could also benefit the NOCF condition). Another possibility is that, for Low-Ed learners, CF is not necessarily beneficial, and that processes of proceduralization are more valuable (i.e. repetition drills) through structured input (cf. processing instruction, VanPatten 2004).

In addition to differences in education backgrounds, the Low- Mid- and High-Ed groups were also found to differ in their length of residence in the Netherlands, the length of Dutch instruction, and computer use. The interlanguage of learners who have been staying in the Netherlands for longer, or have been studying Dutch longer, may have become more resistant to change (cf. fossilization Han & Odlin

---

17 Though none of the participants said they had noticed that the practice targeted V2. Responses on what the practice targeted ranged from listening practice (mostly in the NOCF condition) to speaking, to word-order.
2006). As a result, it may be the case that more CF, or more productions of correct utterances (i.e. longer treatment) are necessary to acquire V2.

Several limitations in this study have to be taken into account. Though we expended a great deal of effort to recruit low educated learners, the number represented in this study is still limited. The difficulty in recruiting low educated learners is one of the reasons why relatively few studies investigate this population (Young-Scholten 2013). We feel that our findings clearly indicate that there are differences between education levels, that results pertaining to one education level cannot be generalized to other education levels, and that this should be further researched with more participants. In addition, the proficiency level of the learners in this study was A1 or A2, on account of the class that they were in (i.e. this was determined by the institutes where they were learning Dutch). We assessed learner proficiency in the pre-test, but it may be necessary to include more detailed pre-test assessments to have more insight into the differences in L2 proficiency of learners with different education backgrounds at the start of speaking practice. Furthermore, in this study we included learners of all L1 backgrounds and ages. A follow-up study could control for L1 and age to investigate education effect more closely. In future research, we suggest that it is important that the education level is reported and its influence analysed, in addition to current practice in CF studies that reports and investigates L1 and age descriptives.

To conclude, the current study has presented findings that L2 practice and CF were not equally effective for L2 learners with different education backgrounds. More research is needed for a better understanding of the effectiveness of practice with or without CF for low educated learners.
Chapter 6: Discussion and Conclusions

1. Introduction

In this thesis, we pursued three main research goals: studying the role of corrective feedback (CF) in learning oral second language (L2) grammar, investigating the influence of individual characteristics on CF effectiveness, and examining the learning process in addition to learning outcomes. Based on studies that showed the necessity of more controlled experiment environments for research on L2 spoken production, Chapter 2 argued for the use of a computer assisted language learning (CALL) system with automatic speech recognition (ASR) to provide optimal conditions for oral practice and CF and to investigate the role of CF. Therefore, developing such a system was an important step to be able to address our research goals. This resulted in the system that we call GREET. In this general discussion, we first deal with the performance of GREET in our experimental environment (6.2), before discussing the three main research goals in order. The Chapters 3, 4, and 5 all address the main goals, to various degrees. Their collective findings are summarized and discussed in sections 6.3 to 6.5. In Section 6.6 we draw general conclusions. Section 6.7 addresses the limitations of the present studies, while the concluding Section 6.8 presents recommendations and perspectives for future research.

2. Prerequisite methodological innovations

As discussed in Chapter 2, pursuing our research goals required an innovative system that provides automatic and consistent CF on oral performance. This can be achieved through the use of ASR in a CALL environment. However, to ensure accuracy of CF, it was necessary to design the content in such a way that the ASR could reliably provide accurate CF to low proficient L2 speakers with varying first language (L1) backgrounds. This required an appropriate linguistic feature on which to provide CF, and the design of sufficiently restricted tasks that elicited pedagogically relevant output. In this section, we first discuss the development of GREET and the experimental environment and then move on to the evaluation.

The ASR for L2 learning employed in GREET had been developed at our department, and had been shown to be able to accurately and reliably process L2 speech, if the limitations of the ASR were taken into account in the design (van Doremalen, Cucchiarini & Strik 2010; see also Strik, Colpaert, van Doremalen & Cucchiarini 2013). As the linguistic target in our study we selected the grammatical feature of verb second (V2), an important part of Dutch word order. With this feature, exercises on word order were designed in a dialogue format in which learners had to form sentences in reply to questions. By instructing the learners to use only the words given on the screen, the ASR could detect the order of words in
The Role of Automatic CF in Learning L2 Grammar

The sentence, and provide CF by comparing the learner’s output with all possible word order sequences. In this way, V2 errors, as incorrect placement of the verb, could be automatically detected. The accuracy of the CF on V2 turned out to be sufficiently high in a pilot experiment (Bodnar, Penning de Vries, Cucchiarini, Strik & van Hout 2011). The target feature was also appropriate since V2 was not noticed as the target under study by the learners (approximately 30% of sentences in the treatment were target sentences, the other sentences did not address any specific feature). The learners thought they were practicing with word order in general (as indicated by them on the post-test questionnaire). The initial proficiency of the learners regarding V2 was variable and most of them could benefit from practicing grammatical accuracy. This indicates that the feature was appropriate with respect to their L2 developmental readiness. The automatic CF on spoken production of V2 was found to have an impact on the learners’ proficiency in the duration of 90 minutes of practice. This need not have been the case, as CF effect differs with the grammatical feature that is targeted (cf. Nagata 1996; R. Ellis 2007; Li 2013). In fact, for low educated learners, the CF type was ineffective for improving V2 proficiency (see Section 6.4 for further discussion).

For maximal consistency, the experiment was designed to be as independent from the researcher as possible: after learners logged in to the website, they were guided with the help of tutorial pages and trial exercises through the questionnaires, proficiency tests and GREET treatment. Streamlining the experiment in this way was realized because the first pilot had revealed the necessity of tutorials. The tutorials were piloted once and used for the experiment presented in Chapter 3. Before the experiments in Chapter 4 and 5, some additional changes were made to the tutorials following pilot experiments with low educated learners.

The behaviour of the learners during the full experiment session was recorded from the moment they logged in to the website. Though this is clearly a benefit for research purposes, it also required decisions on which data on learner behaviour was informative, a well-known side effect of ‘big data’. The logging module was further developed between the experiments. Currently, several measures of learner behaviour can be tracked real-time, which indicates the potential of this approach for future research with logged data such as adaptation and personalization. However, first more research is needed to determine which characteristics of learner behaviour indicate effective practice (i.e. proficiency gain), and which characteristics ask for adaptive exercises and changes in feedback strategies (see also Heift & Schulze 2007 for a discussion on learner modelling).

The evaluation of the system is most explicit in Chapter 3, where improvements were also suggested (and subsequently implemented), and implicit in Chapter 4 and 5. The final version of the system works reliably (given a stable internet connection) enabling independent and individual practice. An important point in the evaluation was that learners could improve their proficiency through practice with the system. The automatic CF, and the design of the exercise, were found to be beneficial for L2 learning. In some cases, practice was not effective (i.e. with the low educated
learners, see Section 6.4). Further research is necessary to improve learning conditions for these learners.

A second important point is the learners’ evaluation of practice with the system. The restrictions on task design as imposed by ASR requirements could have had a negative impact on learner enjoyment, but this was not the case as overall the learners, even those who did not receive CF, evaluated practice with the system positively. The CF condition was generally preferred, showing that automatic CF was perceived to be useful.

A third point for evaluation of the system is whether L2 practice with the system yielded valid information for CF research. As we have shown in Chapters 3, 4, and 5, the experiment environment was controlled and successfully recorded information about the learners in proficiency tests, questionnaires, and log data. These sources could be used to cross-validate findings, and provide us with a more comprehensive analysis of the language learning process.

Finally, with respect to the evaluation of the ASR and the treatment design, the system was shown to successfully cope with a large diversity of L1 backgrounds. In total, 244 learners practiced in some form with the GREET system. 34 learners participated in pilot experiments and 128 learners were included in the research presented in this thesis. In total, there were 32 different L1s in this sample (most common L1s: Spanish, Russian, Arabic; main L2: English).

The GREET system is innovative for several reasons. With respect to pedagogical applications, it allows for skill-specific practice in the spoken modality, which is an important requirement for improving L2 spoken proficiency (DeKeyser 2007). For L2 learners, there are usually only limited opportunities to practice speaking. Practice time in CALL with ASR is potentially unlimited, thereby offering an outstanding practice environment, in particular as practice in CALL is found to cause less anxiety in learners than speaking in class (Warschauer 1996). Sheen (2008) found that anxiety reduces the effectiveness of language instruction.

At the start of our project, there were virtually no systems that provided automatic CF on grammar (Bodnar, Cucchiarini, & Strik 2010), whereas grammatical accuracy is an important part of spoken proficiency. The GREET system was developed to fill that gap. For research possibilities, the GREET system provided individual practice successfully, while system-learner interactions were logged for later analyses and development. This provides the researcher with information on learner behaviour in controlled L2 learning. This data can subsequently be related to other sources of data acquired in the experiment, for triangulation (e.g. proficiency; learner behaviour; questionnaire). Clearly, other data could be elicited from the learner, e.g. explicitly through measures of situated motivation to link to learner behaviour (e.g. Bodnar, Cucchiarini, Penning de Vries, Strik & van Hout (forthcoming)) - or implicitly by allowing learners to toggle CF on or off, or select a CF type. In a setup such as GREET, these data could be combined with proficiency data to provide a comprehensive view of the L2 learning experience.
A final innovative feature of the GREET system is that all the speech produced by the participants in the experiments has been recorded, analysed by the ASR and stored in a database, and is accompanied by speaker metadata (age, gender, proficiency level, educational level, languages spoken, length of residence, length of L2 instruction, and frequency of computer and language use). An important advantage of speech material collected through an ASR-based CALL system is that it comes with all the relevant information for further processing. By compiling the proficiency test data, the personal data, treatment and interaction data, the recorded speech is contextualized in detail, which is highly useful for research on language learning from various perspectives. In addition, it can contribute to developing new, improved language learning systems (see Cucchiarini, Bodnar, Penning de Vries, van Hout & Strik 2014 for a detailed discussion).

3. The role of CF in learning oral L2 grammar

In Chapter 3, we saw that the participants improved their V2 proficiency as a result of GREET practice, in the CF and the NOCF condition. The improvement was found to be similar in the two conditions: CF was effective, but did not outperform spoken practice without CF. However, learners in this experiment were relatively high proficient, and closer investigation showed that the NOCF group participants with a low pre-test proficiency did not improve their accuracy on the target structure. This suggests that while learners with higher proficiency levels may learn without CF, those with lower proficiency may require CF to improve their grammatical accuracy. Chapter 4 followed up on this suggestion, and learners were recruited at a lower proficiency level. In an analysis that excluded ceiling participants, learners in both conditions improved their V2 proficiency. On the measure of spoken production (DCT), learners in the CF condition outperformed those in the NOCF condition, thus showing that spoken practice with CF was more effective than spoken practice alone. Chapter 5 investigated whether these findings apply to learners with lower educational background. The participants had various educational backgrounds (primary school to PhD degree), but were at the same proficiency level as in Chapter 4 (excluding again ceiling participants). The results showed that learners in both conditions improved on the measure of spoken production (DCT). There was no additional effect of CF, but this additional effect was found when the low educated participants were excluded (see Section (6.4) for further discussion).

In Chapter 3 and 4, learners rated practice with GREET positively, and learners were found to prefer the CF condition. In Chapter 5, practice was rated positively, but no difference between the conditions was found. The higher rating for the CF condition lined up with more efficient practice in the CF condition in Chapter 4.

Overall, learners improved their V2 proficiency as a result of GREET practice. Both the CF and the NOCF condition resulted in proficiency gain, and were evaluated positively. Providing automatic CF on spoken practice was found to be
more effective than spoken practice without CF for learners with a low or medium proficiency level, but this applied only to high-educated learners.

How do these findings inform CF research? First of all, we find that CF does not have a negative effect on L2 learning. In none of the studies do learners in the CF condition perform worse than learners in the NOCF condition (cf. Krashen 1982; Truscott 1996). Moreover, as the CF condition was often rated more positively than the NOCF condition, learners seem to appreciate CF.

Learners in the CF condition improve their proficiency, but this is also true for learners in the NOCF condition. Practice with our program is beneficial for L2 spoken proficiency, and its effectiveness is not purely related to the presence of CF. The effectiveness of the CF and NOCF condition seem to be related to different cognitive processes and the degree of their involvement during practice.

In the NOCF condition, learners produced an utterance with the knowledge that they would not receive immediate CF, but that their attempts would be judged and stored by the system. Production of a spoken utterance (with the target feature) is beneficial (a) from the point of view of comprehensible output (Swain 1985), which may cause learners to notice errors in their L2 speech (cf. Schmidt 1995; Kartchava & Ammar 2013), (b) for automatization of explicit knowledge (i.e. applying the V2 rule in output) (De Bot 1996), and (c) from a skill-specific practice point of view (DeKeyser 2007).

We find that learners in the NOCF condition practice roughly the same amount of questions during practice as the learners in the CF condition, but also generally make only one attempt. This implies that the learners spend more time preparing their (first) answer. They are also aware that they will not receive CF from the system, and that they must construct and judge the answer themselves. It seems plausible to assume that practice in this condition requires more cognitive involvement. This is likely to be a beneficial condition for L2 practice (cf. Robinson 2003), and may contribute to the three processes mentioned above. The treatment also provides the learners with task-essential practice on the target structure, and this by itself is effective for learning (VanPatten 1994; Sanz & Morgan-Short 2004). In Chapter 3 we argued that the NOCF condition's effectiveness could be in part related to the fact that learners feel a sense of interaction, even without immediate CF, because a sense of interaction and monitoring by the system is enough to stimulate learners to make the most of the input (Chapelle 1998). Ultimately, though, the effectiveness of practice in the NOCF condition relies on the learners' efforts to produce a grammatically accurate sentence, and their skills to accomplish that, or to notice their own errors.

In the CF condition, learners produced an utterance with the knowledge that they would receive immediate CF. In comparison to the NOCF condition, learners in

---

18 Note that (b) and (c) apply for grammatically correct application of the V2 feature for learning the V2 target feature. Furthermore, it is not unlikely that, as a result of practice, learners also improved on other features of L2 proficiency (morphology, listening or reading skills, lexical knowledge, pronunciation), but these are not measured.
the CF condition may have been less concerned with accuracy in their attempts, as they know the system will provide assistance (the data shows that learners in the CF condition produce attempts faster than the NOCF participants, discussed in Section 6.5). The learners produce output (containing V2) and receive explicit information on grammaticality, and after two or three incorrect attempts they are given input regarding the target structure. Similar processes (a-c) as mentioned above for the NOCF condition take place in the CF condition, but are stimulated. Since the learners are prompted to retry incorrect answers, they are pushed to produce more output: modified output, and more grammatically correct output (since the CF condition forces a correct utterance before it allows the learner to proceed). In addition, noticing the error is facilitated through the explicit CF, and learners are sometimes (i.e. after several errors on the same question) provided with input regarding the correct form (cf. VanPatten 2004).

When comparing the processes in the two conditions, the improvement in the NOCF condition can be explained as it is dependent on the learner's effort to produce a grammatical utterance, and on their proficiency to be able to notice errors or produce grammatically accurate utterances. Under certain circumstances, then, it can be as effective as the CF condition. This is similar to the results of the separate experiments.

In specific circumstances, which were related to individual differences, CF was found to have additional learning benefit over spoken practice. By controlling for proficiency level and education level, CF outperformed the NOCF condition. For learners with a low education level, both types of practice, and thus also CF, were ineffective. As our discussion of L2 research in Chapter 1 and 2 has shown, there is much variety between findings in CF research. The findings in this controlled study provides insights into this variability. In an experimental setup that was overall highly constant, the findings about the effect of CF vary on account of individual differences.

With respect to the evaluation of practice, the learners felt that practice with and without CF was enjoyable and beneficial for L2 development. When there was a preference for the CF condition, this did not necessarily coincide with more proficiency gain in the CF condition. We can relate this to the findings that learners indicate preferences for receiving CF (Lyster, Saito & Sato 2013; though see Loewen et al. 2009 for a contrary finding), and that they feel they practice more effectively with CF (Van der Linden 1993; Jean & Simard 2011). A CALL system is beneficial in this way, as teachers often do not want to provide CF as often as learners would like (Jean & Simard 2011). Learners also rated the NOCF condition positively: we might have expected a negative influence of the learners being disengaged through the absence of CF (cf. Shulz 1996). Apparently, the task was perceived as useful even without CF. Influentially, Truscott (1999) and Krashen (1982) argued against an effect of CF for the reason that it caused anxiety (which has shown to impede L2 learning (Sheen 2008)). Although CF in a CALL system is less likely to cause anxiety (cf. Warschauer 1996 on L2 production in written chat.
conversations), it is an important observation that the learners in our experiments did not indicate any dislike of the CF.

4. How do Individual differences influence CF effectiveness?

The literature overview in Chapter 2 (and Chapter 4) shows that CF research has not addressed individual differences in sufficient detail (cf. R. Ellis 2010; Li 2010). With GREET, it is possible to inspect individual differences and their influence on L2 spoken practice. In our study we specifically addressed the role of learner proficiency and education background in L2 practice, with and without CF.

In Chapter 3 there was no difference in learning gain between the two conditions, but inspection of individual trajectories suggested that the initial proficiency of the learners was related to the effectiveness of the experimental condition. With a lower proficient learner sample in Chapter 4 this hypothesis was inspected more closely, and confirmed. The CF condition outperformed the NOCF condition, for low and medium proficient learners. In Chapter 5, learners with lower levels of education worked with the GREET system. Analysing within the same proficiency level as Chapter 4, we found that learners with a low education level (primary school) did not benefit from practice in either condition (CF nor NOCF). Learners with higher levels of education improved in both conditions, and there was a trend for the mid- and high educated learners to perform better in the CF condition, in line with Chapter 4.

These results demonstrate that individual differences influence the effectiveness of CF. The learners’ proficiency influences what they can notice in the input (Hanaoka 2007; R. Ellis 2000): for instance, a low proficient learner will tend to focus on the lexical items, and is more likely to overlook grammatical features. However, when CF is provided, the learner’s attention is drawn to specific errors (e.g. the grammatical features). In addition, in the case of prompt CF as provided by our system, the learners are pushed to reformulate, and thereby consciously or subconsciously apply the correct grammatical rule. This may explain the difference between the effect of the NOCF and the CF condition: low proficient learners in the NOCF condition in our study may have been focusing on aspects of e.g. pronunciation instead of, or in addition to, the target feature. Moreover, as discussed in (6.3.) above, low proficient learners may not have sufficient L2 knowledge to notice a V2 error, or to produce a correct instance of V2. Ammar & Spada (2006) show that the type of CF that is effective changes with the learner’s proficiency. They found that low proficient learners benefitted more from explicit CF than from recasts. In our study, all learners received explicit CF (prompts), which is in line with Ammar & Spada’s findings. If we develop more content that has several levels of L2 difficulty, it will be interesting in future research to compare how the effective type of CF changes with proficiency.
However, at a similar proficiency level, learners with lower education levels did not benefit from practice. For both conditions, it may be the case that learners did not notice the feature, or did not pick up a regularity in their errors as relating to the same feature, or did not have sufficient time/opportunities to practice. In the CF condition, these learners apparently could not benefit from CF. In Chapter 5, we assumed that the reasons for the difference in performance between high educated and low educated learners (in terms of proficiency gain) may have been their lower developed cognitive and metacognitive skills, or, perhaps a lower concern with L2 accuracy (overall, or simply during practice). A different type of CF might have been more beneficial. For instance Bigelow, Hansen, Delmas & Tarone (2006) showed that low literate learners and low educated learners could not benefit from recasts because they did not notice the corrective intent of the CF. The CF in our study is explicit, but does not specify the error, nor provides a rule. It is therefore possible that the low educated learners did not notice the corrective intent in our study. Sagarra & Abbuhl (2013) find that enhancing oral recasts by stressing the error improved learning results (though not for written recasts). Perhaps this enhancement of the CF message is particularly necessary for low educated learners. It may also be the case that low educated learners need more instances to practice for improvement to occur. In this way, learning could also be implicit, through (relatively) high frequency of V2 in the input (cf. N.Ellis 2002, 2006). Of course, also a different type of exercise may be necessary, for instance where the learners are first explained the rule, before starting spoken practice. Clearly, however, other types of CF and practice should be researched for effectiveness with low educated learners.

The results of the experiments combined provide an articulated picture of CF effectiveness and its relationship to learner variables. The role of CF and its contribution to language learning is seen to vary depending on the proficiency level of the learner on the specific, targeted grammatical structure. In addition, if the level of proficiency improves as a result of practice, the role of CF is likely to vary accordingly. The role of CF also cannot be simply generalized to account for learners of all educational levels. To understand in more detail the differences in outcome of spoken practice with and without CF, it is necessary to understand how the learners interacted with practice and CF.

5. What is the impact of CF on learner behaviour?

Perspectives on L2 learning such as sociocultural theory, interactionist theory, and dynamic systems theory (Aljaafreh & Lantolf 1994; Mackey & Gass 2006; De Bot, Lowie & Verspoor 2007), and in addition the disagreements in SLA research on beneficial learner behaviour (e.g. the role of uptake) emphasise the importance of analysing the learning process (Chapelle 2009b). As our study shows, the effects of CF differ according to individual differences. Outcome measures do not provide information on learner interactions with CF, and thus do not inform us why CF may
be more or less effective for individuals. GREET logged the data of the learners, which allowed insight into how their practice sessions proceeded.

In Chapter 3, inspection of learner behaviour in the two conditions revealed significantly different practice sessions, even though learners’ proficiency had improved equally. The groups practiced an equal number of questions, but the CF group produced significantly more spoken output, and repaired their errors significantly more often.

Though we improved the logging module of the GREET system to record more learner behaviour data (e.g. time spent per page, time reading questions, etc.) for the experiment in Chapter 4, analyses revealed that the relevant features remained the number of questions practiced, and number of (correct) attempts. In Chapter 4, there were also clear differences in learner behaviour between the two conditions. Here, however, the CF condition was found to outperform the NOCF condition. Analyses showed that the number of questions practiced was equal between conditions, and that the CF condition produced more spoken output overall, and more correct instances of V2. 19

Learner behaviour as given by the total number of questions or attempts practiced in 90 minutes was related to learner proficiency (i.e. a higher proficient learner could practice more questions), but did not correlate with proficiency gain. In the CF condition, learners required fewer instances of CF and attempts to provide the correct answer as they progressed through practice. Evidence of learning could be found in both conditions, as the learners improved their ratio of correct answers over time. As a result, we were able to link proficiency gain in the post-tests to characteristics of learner behaviour.

In Chapter 5 we inspected learner behaviour to increase our understanding of why the low educated learners did not improve as a result of practice. A salient difference between the educational levels was the number of correct sentences produced, which was lower for low educated learners. Similar to Chapter 4, we analysed the change over time in the learner sessions. Here we found that learner behaviour varied according to educational level. High educated learners improved their proficiency, used fewer instances of CF over time, and produced a higher ratio of correct answers over time; Medium educated learners improved their proficiency, but their learner behaviour did not change significantly. Low educated learners did not improve their proficiency, and their learner behaviour did not change significantly.

In the discussions in Chapter 3, 4, and 5, learner behaviour is regarded as indicative of cognitive processes. As Hegelheimer and Chapelle (2000) discuss, only indirect inferences can be made about the processes that occur during practice. Through comparison and correlation with proficiency test data, we find that it is possible to distil relevant learner behaviour.

---

19 The number of repairs from Chapter 3 was broken down into two measures: the attempts per question and the number of correct utterances.
The number of questions that were practiced were relatively similar between the conditions. This is relevant from the point of view of input processing and task-essential practice (Sanz & Morgan-Short 2004; VanPatten 2004). As we have argued earlier (6.3), this may have been effective for practice in both conditions, though perhaps more so in the NOCF condition, as learners were more cognitively involved.

The clearest difference between the experiment conditions is that the CF condition produces more attempts, and more attempts per question than the NOCF condition. This is the result of CF, which requires learners to retry after an error. The number of attempts in the CF condition should be seen as both (a) more produced output (relevant for production of a V2 sentence (Swain 1985; De Bot 1996) and skill-specific practice (DeKeyser 2007)), and (b) as received instances of CF (facilitates noticing the error (Schmidt 1995), or confirmation of correct), and for providing (modified) input (Long 1996; Gass 1997), whereas in the NOCF condition, the number of attempts only indicates (a) produced output.

Based on the findings in the three experiments, we assume that produced output per se may not be the important measure, but rather the amount of produced output that is grammatically correct. Whether this is more important than the instances of CF received is difficult to discern in our data, but certainly the CF stimulates the learner to produce more grammatically correct utterances. In addition, the CF seems to motivate the learners to produce more output (which may be beneficial also for non-target skills such as pronunciation).

In analyses of changes in learner behaviour over time, effectiveness of practice (in terms of proficiency gain) was found to co-occur with a decrease in the number of attempts that a learner needed to correctly answer a question. As a result, we argue that characteristics of learner behaviour can be employed to signal learning. For future use, this measure could be used to determine if the CF type was effective: if the number of attempts per question does not decrease, a different CF type can be applied, or it can be seen as an indication that the current difficulty level of practice is too hard. Using this measure would indicate that, in Chapter 5, the number of attempts per question stays high for the low educated learners, and practice should be adapted. Through trial of different instructional techniques, an effective method could be determined. Learning could also be detected in the learner behaviour in the NOCF condition, as proficiency gain coincided with the accuracy on answering questions improving. This measure can be used to determine when a learner may not be practicing effectively, and (a type of) CF could improve performance.

6. General conclusions

From the findings in this study, we can conclude that CALL systems with ASR are an important innovation for SLA research that allows for practicing oral skills with CF under controlled experimental conditions, with detailed logging of learner behaviour. However, design of the CALL exercises is crucial for the ASR technology to be effective, and therefore requires an interdisciplinary approach (SLA, pedagogy, interface design, CALL, speech technology), and careful
evaluation. Practice with our system was appreciated by learners and turned out to improve language proficiency regarding a specific target feature. The study shows that claims about CF effectiveness in L2 learning require nuance. CF is found to be more effective than spoken practice without CF, but this is in interaction with specific learner variables.

Individual differences between learners mediate CF effectiveness. The learners’ L2 proficiency is an important factor. Since proficiency is a broad indicator, proficiency should be specified according to the target feature. Learners at low to intermediate level of proficiency benefit more from CF than from spoken practice alone. However, we find that the results of high educated learners differ from those of low educated learners. Low educated learners did not benefit from spoken practice, with or without CF. CF effectiveness, therefore, does not generalize to learners with different education levels.

Learner behaviour during practice with a CALL system provides valuable information. For research purposes, logging of learner behaviour provides information about individual variation in practice, and quantifies amounts of L2 input received and output produced, and how these change over time. This information allows for inferences about effective L2 practice and can thus inform SLA research on processes during interaction with CF that contribute to learning, and CALL design on possibilities for adaptive systems.

7. Limitations

In the research reported on in this thesis, we can identify several limitations which we discuss here.

Sample size

The first, and perhaps the most common limitation, is the number of participants. L2 learners of Dutch are a relatively small group, and recruitment of L2 learners of Dutch (especially low educated learners) was difficult and time consuming. Though many SLA studies have smaller participant samples than ours, a larger number of participants would have made our study more powerful (statistically). A great benefit of research on oral production with a CALL system is that it allows for accumulating individual participants over longer time and at different locations. An example is shown in Chapter 5. GREET practice is consistent, which allows for a long period to add participants and to use diverse samples of participants. For studies on ‘smaller languages’ like Dutch, our setup seems very well suited to gather learners from multiple sources. With more participants, the effects can be determined with more statistical power, and in addition, this will allow for identification of subgroups. As we have done in Chapter 4 and 5, through correlational analysis of personal information gathered in a pre-questionnaire, learning outcome, post-questionnaire evaluation and learner data, relevant characteristics of learners or behaviour can be identified. For instance, the L1 is
likely to influence learner behaviour (Sheen 2004) or outcome (Bohnacker 2006; Ringbom 1992), but also age, gender, age of arrival have been shown to influence learning and the effects of these variables could be compared. With more participants, the influence of individual differences can be studied in more detail. Ideally, the system would be part of a Dutch L2 course at various institutes, which would allow for a continuing flow of participant data.

Assessment of proficiency

In CF studies, measurement of the effect of CF on L2 proficiency is clearly crucial. The assessment of proficiency is subject of ongoing study (cf. Rebuschat 2013; Blom & Unsworth 2010), especially with respect to the measurement of implicit L2 knowledge. The difference between explicit and implicit L2 knowledge plays an important role in the CF debate. Especially for L2 proficiency tests, responses are based, to varying degrees, on both implicit and (speeded) explicit knowledge.

In our study, much care was taken in the development and selection of the proficiency tests, but some limitations should be pointed out. We used two proficiency tests (cf. Hulstijn 1997; Norris & Ortega 2003; Mackey 2006), which were selected based on R. Ellis’s (2005) psychometric study. The tests were cross-validated through correlation analyses of these tests with each other and to language behaviour during practice (i.e. logically, learners with higher scores on proficiency tests should perform better during practice). It is important to use multiple tests to assess L2 proficiency, especially because the type of task used for measurement influences the effect sizes of results (Mackey & Goo 2007), with closed-ended response tasks yielding the largest mean effect, and acceptability and selected response measures the smallest.

We used the DCT task as an oral production task (R. Ellis 2005), which classifies as a constrained response test. Free open response tasks are considered by some researchers to be better indicators of implicit knowledge (cf. Spada & Tomita 2010), but since we targeted a specific feature, this was not feasible (i.e. due to the possibility of avoidance: in Dutch, our target feature could easily be avoided). We included a time limit (30s) to ensure that learners did not ponder sentences too long, but in some instances, learners corrected themselves (i.e. they could retry within the time limit). So it may seem that in some cases the learners drew on their (speeded) explicit knowledge to respond (note, however, that explicit knowledge may assist in the forming of implicit knowledge, so even if CF has an impact on explicit knowledge, this does not argue against the fact that CF can impact implicit knowledge, be it indirectly).

The timed GJT is a frequently used instrument in SLA. However, several issues are mentioned regarding this measure: the learners may be influenced in their judgment by the semantic content of the sentence (to control for this, we paired grammatical and ungrammatical sentences that were identical save for the error). Learners may be grading acceptability and not strict grammaticality, e.g. sentences that may be ungrammatical can still be perfectly understandable (Bard, Robertson &
Sorace 1996). For our particular target feature that is a problem, since violation of V2 does not affect meaning. To address this problem we included a gradient scale in which learners could indicate ‘degrees of grammaticality’. However, this has been argued to stimulate learners to use metalinguistic knowledge (Bader & Häussler 2010), and in addition, we found that the instruction for the test was complicated for the learners: for them a yes/no response seemed more intuitive. In our results, we did not find evidence that learners used the gradient, and usually answered at the extremes of the bar (i.e. a binary response). Analysis of the gradient GJT scores did not yield additional information.

In the GJT, the setting of the time limit is an important factor. With unlimited time, learners rely on explicit knowledge to respond (cf. R. Ellis 2005). However, in the GJT there is the possibility of using non-linguistic information to provide an answer (i.e. guessing). The problem of guessing is increased by making use of the time pressure. According to Loewen (2009) 10 seconds on the GJT is the upper limit for use of implicit knowledge. In our study we set the limit to 12 seconds on the basis of pilot tests with low and high educated learners. To ensure that learners felt a sense of time pressure, the GJT interface had a time bar that ran out. We found that this was effective in giving learners the sense of pressure, as many learners (of mixed proficiency) remarked afterward that the test was fast. Another source of interference in the GJT is fatigue. We inspected the results for signs of fatigue, but did not find a significant influence. In our study, we found that the GJT was informative for high educated learners, but not for lower educated learners, as their scores were around chance level. Reading skills, or concentration, may have been a cause. To reduce reliance on reading skills, it is possible that an aural GJT may be better suited to compare low and high educated learners, which will be interesting to research in future experiments.

Use of the logged data on learner behaviour

As a result of logging learner behaviour, there is a great deal of information available about the practice sessions. Our analyses have focused on a linguistic interpretation of this behaviour, but clearly more research is possible on what information in the logs is indicative of in-/effective CALL design for the user interface and tutorials.

Design of the GREET treatment

The design of the program has an effect on learners. For example, most learners enjoyed the videos in GREET, but there were also some who did not like them. Also the kind of exercise, and the type of CF may appeal to some learners, and not to others. For experimental control, the interactions in GREET were restricted. Clearly, the engagement of a learner with the learning content is an important factor in determining practice effectiveness (cf. R. Ellis 2010). In this study, learners generally appreciated practice by rating it positively in the questionnaire. However,
it is likely that if learners had had more options to indicate preferences (e.g. for type of CF, or type of videos) during practice, the effectiveness of practice could have been higher. In the current study we did not investigate in further detail the affective learner variables in practice with CALL, although this is an interesting possibility (Bodnar, Cucchiarini, Strik & van Hout 2014).

Design of CF in GREET

A factor influencing CF effectiveness is CF type. In this study we decided to analyse one type in depth, instead of comparing CF types. Clearly, however, we can expect that a different CF type, or even different operationalizations of CF might have yielded different results. Fortunately, the CF type is concretely defined in the CALL system, which allows for cross-study comparison. In our setup, comparison of CF types is an attractive future possibility, or allowing learners to toggle CF on or off, or selecting the type of CF, which in turn could inform us on learner preference (cf. Hegelheimer & Tower 2004). With respect to the type of CF that we provided, we suggest a possible improvement, which is the possibility of repeating the correct utterance. Currently, learners may struggle to provide a correct answer. If they really do not know (guess), they may be surprised when they finally answer correctly. The system does not allow learners to retry a correct answer (in the CF condition). As a result, learners may sometimes miss the opportunity to understand the correct answer. From a procedural and noticing point of view, it may be beneficial for learners to be able to repeat a correct answer – in a way for them to make sure they understand.

8. Future research

As we have seen, a CALL system with ASR such as GREET is effective for research on oral CF, and for improving L2 proficiency. In our view, these systems hold great promise for future research. In SLA it is becoming increasingly clear that the effect of CF is mediated and moderated by many variables, which requires a new approach to CF research. In our study, we have investigated one target feature and investigated in detail how learner variables interact with CF on learning that feature. In a similar setup with GREET, it is also possible to make comparisons of learning with different types of CF (e.g. recasts and metalinguistic CF), or with different linguistic targets. Since the system is otherwise consistent, the influence of CF type could be methodically investigated. Moreover, if the CALL system can register the learners affective- or motivational state, this can be related to CF (type) effectiveness, and inform learner behaviour (Bodnar et al. 2014; see for an example Bodnar, Cucchiarini, Penning de Vries, Strik & van Hout (forthcoming)).

By adding more instructional content to the system (or such a system), longitudinal studies in CALL are possible (see Heift 2010 for an example). Longitudinal studies allow for other related research questions to be examined: the
within-subjects interactions of linguistic targets and CF types, and how these change over time - and how this differs between subjects. Clearly, then, this would provide opportunities to test a system that automatically adapts to the learner by analysing learner behaviour. Adapting the difficulty level to the level of education and level of proficiency of participants offers new opportunities to research CF types (effectiveness for learning new knowledge, or strengthening learned knowledge), as we mentioned in Chapter 4. In addition, this would allow investigation into the notion of the zone of proximal development (ZPD, Vygotsky 1978). In our study, we have two measures that were seen to change over time and relate to L2 learning (ratio of correct answers, and attempts per question) These measures provided information about the relative difficulty of the task for the learner. A system could adapt the difficulty level of practice to keep learners in the optimal learning zone: for instance that, for a given time frame, they need on average 1.5 attempts to answer, or that 75% of their answers are correct. Setting these levels could inform us on the nature of the ZPD, and make treatment maximally efficient (and perhaps also maximally engaging). CF types could be compared in their efficiency in helping learners move to the next difficulty level. In addition, it will be interesting to research the relation of CF preference, CF perceived effectiveness and CF effectiveness. Through close monitoring of learner behaviour and in-task reports of motivation, or through retrospective reports, the engagement of the learner could be related to CF effect.

Of key importance for the effectiveness of a CALL system is the learner’s motivation, or the willingness of the learner to continue practicing. A promising new development in the field of CALL is the growing interest in serious games. As seen in the study by Cornillie (2014), practice with these systems can provide effective CF on a language learning experience where the learner is immersed in the game environment. A combination of serious games and spoken practice as employed in GREET would allow to study CF in a context that is more meaning-based and probably engaging.

Other interesting directions for future research and CALL development concern relating CALL practice more closely to the language learning goals of the L2 learner, which should contribute to increasing motivation (Colpaert 2010). This would of course require the use of appropriate tools for eliciting those goals, and the development of content to address these goals. This is an exciting new direction for CALL research, and GREET could be employed in such research (Bodnar et al. (forthcoming)).

A current bottleneck for longitudinal studies with CALL is the creation of appropriate L2 content. The internet contains vast resources, such as film clips (with varying levels of meta-information, such as transcriptions), which presents us with the possibility of automatic generation of materials (cf. Presson, Davy & MacWhinney 2013). It is clear that this holds great promise for expanding the possibilities for pedagogical CALL applications, and interrelated, for L2 research with CALL. In addition, experiments through the internet, such as sometimes already run through crowd sourcing (Rayner & Tsourakis, 2013), can boost the number of participants: more learners can be approached (even internationally) and
learners may be more inclined to participate, as they can simply work from home. For experimental control, this poses problems (e.g. learners may be discussing tasks with others, or not be focused on the task with a television in the background), and these need to be addressed in the experimental setup. The analysis of learner behaviour may be of use here, as the program can provide feedback to the participant (‘hey, make sure you answer your questions in time!’) or notify the researcher of irregular behaviour.

With respect to logging behaviour, the GREET system could be equipped with an eye tracking system (see for an example of a study using eye tracking, Smith 2012). These systems are becoming increasingly available, and could easily be attached to an experiment computer. For our experiment, this may add an extra layer of learner information: by studying eye movement, we could infer how they read the messages and the blocks on the screen, which could be analysed for signs of noticing the CF. Also for the proficiency tests this adds valuable extra information: especially for the GJT this may prove valuable (Godfroid, Loewen, Jung, Park, Gass & R. Ellis 2014).

The findings of our experiments were obtained in a highly controlled environment. In a classroom, such controlled settings are not possible, and CF will be provided and received in a much more diffuse manner (cf. Lyster et al. 2013). Our findings may therefore not transfer directly to the classroom. On the other hand, we have to consider that CALL and blended learning are becoming increasingly important in L2 pedagogy (Chapelle 2007, 2009b; Warschauer 2007) and the use of more advanced technologies will probably increase their functionalities and applicability. It is therefore necessary to investigate L2 learning in this new context. With the advance of technological possibilities, however, it is necessary to emphasise that the effectiveness of these technologies for L2 learning are dependent on the design in which it is implemented, and that evaluation and development of language learning systems is a crucial step in improving pedagogical design (cf. ‘Educational Engineering’, Colpaert 2014; Chapelle 2001).

Our experiment with the GREET system was performed mostly outside of the class. However, it is easy to see that practice with this system could make it into a language course. Particularly for speaking practice, which requires much repetition and CF on features such as pronunciation, or grammatical features such as articles, V2, or morphology, tutorial CALL systems such as ours can be very valuable for L2 learning. Ideally, in the future, systems such as GREET could be part of L2 learning courses for speaking practice, for learners of various educational backgrounds. The practice data of the learners could be analysed to inform SLA, and the design of the CALL system. In that manner, CALL systems will not only generate new findings for SLA, but also provide data that can be analysed to improve their effectiveness.
References


References


The Role of Automatic CF in Learning L2 Grammar


The Role of Automatic CF in Learning L2 Grammar


The Role of Automatic CF in Learning L2 Grammar


References | 157


The Role of Automatic CF in Learning L2 Grammar
Appendix
### Appendix A:

**Grammaticality judgment task items (Chapter 3)**

<table>
<thead>
<tr>
<th>Id</th>
<th>Sentence (* marks incorrect)</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In de winkel ziet Laura haar vriendin</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>Vandaag zijn de bananen in de aanbieding</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>* In het weiland de koeien eten gras</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>* Vanavond Thomas komt bij ons eten</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>Zolang de trein rijdt mag je niet uitstappen</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>Door drukte op de weg kwam Tim later</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>* Als het blijft regenen dan alle spullen worden nat</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>Op de markt hebben ze nieuwe schoenen voor haar gekocht</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>Elke maandag gaat Lisa’s moeder met de bus naar de stad</td>
<td>B</td>
</tr>
<tr>
<td>10</td>
<td>* In de krant een verhaal staat over de vader van Lotte</td>
<td>A</td>
</tr>
<tr>
<td>11</td>
<td>* Na de ruzie Sofie en Anouk hebben het goed gemaakt</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>Voordat Lucas mag spelen moet hij eerst afwassen</td>
<td>A</td>
</tr>
<tr>
<td>13</td>
<td>Als de regen voorbij is fietst Max naar school</td>
<td>B</td>
</tr>
<tr>
<td>14</td>
<td>* Om vijf voor negen de bus vertrekt naar Rotterdam</td>
<td>A</td>
</tr>
<tr>
<td>15</td>
<td>* Twee keer in de week Daan gaat voetballen</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>Als Koen naar huis gaat, dan neemt hij bus veertien</td>
<td>A</td>
</tr>
<tr>
<td>17</td>
<td>De kat blijft thuis, daar wacht hij op zijn eten</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>* De jongen zit thuis en daar hij wacht op Stefanie</td>
<td>A</td>
</tr>
<tr>
<td>Page</td>
<td>Sentence</td>
<td>Translation</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>24</td>
<td>* Rick koopt een nieuwe tas, dan hij kan boodschappen doen</td>
<td>Rick buys a new bag, then he can get groceries</td>
</tr>
<tr>
<td>27</td>
<td>* Milan kookt soep, hopelijk zijn moeder vindt het lekker</td>
<td>Milan cooks soup, hopefully his mother will like it</td>
</tr>
<tr>
<td>28</td>
<td>* Als je auto gestolen wordt, dan je moet de politie bellen</td>
<td>If your car is stolen, then you should call the police</td>
</tr>
<tr>
<td>29</td>
<td>Hij maakt zijn huis schoon en dan gaat hij tv kijken</td>
<td>He cleans his house, and then he watches tv</td>
</tr>
<tr>
<td>30</td>
<td>De crimineel zit in de cel en daar denkt hij aan vrijheid</td>
<td>The criminal sits in his cell and there he thinks of freedom</td>
</tr>
<tr>
<td>31</td>
<td>* De vrouw koopt een auto en dan ze rijdt naar haar ouders</td>
<td>The woman buys a car and then she drives to her parents</td>
</tr>
<tr>
<td>32</td>
<td>* Nick is klaar met werken en nu hij gaat naar huis</td>
<td>Nick has finished work and now he is going home</td>
</tr>
<tr>
<td>33</td>
<td>Aan hem gaven ze het boek</td>
<td>To him they gave the book</td>
</tr>
<tr>
<td>34</td>
<td>Voor Iris hebben haar vriendinnen een feest gegeven</td>
<td>For Iris her friends organised a party</td>
</tr>
<tr>
<td>35</td>
<td>* Voor de buurman Tom maakt een taart</td>
<td>For the neighbour Tom makes a cake</td>
</tr>
<tr>
<td>36</td>
<td>* Aan zijn hond Mark geeft alleen het lekkerste voer</td>
<td>To his dog Mark only gives the tastiest food</td>
</tr>
<tr>
<td>37</td>
<td>Twee tomaten en een ui hebben we nodig</td>
<td>We need two tomatoes and an onion</td>
</tr>
<tr>
<td>38</td>
<td>Gele borden op de snelweg heb ik nog nooit gezien</td>
<td>I have never seen yellow signs on the motorway</td>
</tr>
<tr>
<td>39</td>
<td>* Een kale man met een fiets Lieke zag op straat</td>
<td>A bald man with a bicycle Lieke saw on the street</td>
</tr>
<tr>
<td>40</td>
<td>* Een spijkerbroek met veel zakken Emma vindt leuk</td>
<td>Jeans with many pockets is what Emma likes</td>
</tr>
</tbody>
</table>
## Appendix B:
### Discourse completion task items (Chapter 3)

<table>
<thead>
<tr>
<th>Id</th>
<th>Context Sentence</th>
<th>Lead-in Sentence</th>
<th>Clue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>In Amsterdam...</td>
<td>veel mensen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>many people</td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td>Bij de bank...</td>
<td>rijke mannen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At the bank...</td>
<td>rich men</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>Onder de paraplu...</td>
<td>hond + droog</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Under the umbrella</td>
<td>dog + dry</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>Om tien uur 's avonds...</td>
<td>naar bed + Linda</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At ten o’clock at night...</td>
<td>to bed + Linda</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>Marloes is geslaagd en nu...</td>
<td>feest + geven party + give</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marloes has passed and now...</td>
<td>party + give</td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td>Dennis zet de tv uit en dan...</td>
<td>Hardlopen jogging</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dennis switches the TV off and then...</td>
<td>jogging</td>
<td></td>
</tr>
<tr>
<td>5a</td>
<td>Patrick gaat met de trein naar Amsterdam.</td>
<td>Daar there</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patrick takes the train to Amsterdam.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5b</td>
<td>Wanneer ga je naar de dokter?</td>
<td>een afspraak an appointment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>When are you going to the doctor?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6a</td>
<td>Wil je vandaag werken?</td>
<td>thuis + blijven home + stay</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Do you want to work today?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6b</td>
<td>Wendy gaat met de trein naar Amsterdam.</td>
<td>de trein + vertrekken the train + leaves</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wendy takes the train to Amsterdam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7a</td>
<td>Wanneer komt Patricia terug naar Nederland?</td>
<td>Patricia Patricia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>When is Patricia coming back to the Netherlands?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7b</td>
<td>Moeder heeft een verjaardagstaart gebakken.</td>
<td>15 kaarsjes 15 candles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mother has baked a birthday cake.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8a</td>
<td>Dit is een hele mooie vijver.</td>
<td>In de vijver...</td>
<td>veel + eenden</td>
</tr>
<tr>
<td></td>
<td><em>This is a very nice pond.</em></td>
<td>In the pond...</td>
<td>many + ducks</td>
</tr>
<tr>
<td>8b</td>
<td>In de boom zit een mooie vogel.</td>
<td>Met zijn camera...</td>
<td>foto + maken</td>
</tr>
<tr>
<td></td>
<td><em>There’s a pretty bird in the tree.</em></td>
<td><em>With his camera...</em></td>
<td><em>picture + make</em></td>
</tr>
<tr>
<td>9a</td>
<td>De weerman op tv zegt:</td>
<td>Morgen...</td>
<td>warm</td>
</tr>
<tr>
<td></td>
<td><em>The weatherman on television says:</em></td>
<td>Tomorrow...</td>
<td>warm</td>
</tr>
<tr>
<td>9b</td>
<td>Het is koud.</td>
<td>Gelukkig...</td>
<td>een warme jas</td>
</tr>
<tr>
<td></td>
<td><em>It’s cold</em></td>
<td>Luckily...</td>
<td>a warm coat</td>
</tr>
<tr>
<td>10a</td>
<td>Ga je mee naar de film?</td>
<td>Nee, vanavond...</td>
<td>studeren</td>
</tr>
<tr>
<td></td>
<td><em>Do you want to go to the movies?</em></td>
<td>No, tonight...</td>
<td>study</td>
</tr>
<tr>
<td>10b</td>
<td>Wat doet u op zaterdag?</td>
<td>Op zaterdag...</td>
<td>de stad + gaan</td>
</tr>
<tr>
<td></td>
<td><em>What will you be doing on Saturday?</em></td>
<td>On Saturday...</td>
<td>the city + go</td>
</tr>
<tr>
<td>11a</td>
<td>Wat doet Kim om twaalf uur?</td>
<td>Om twaalf uur...</td>
<td>boterham + eten</td>
</tr>
<tr>
<td></td>
<td><em>What does Kim do at twelve o’clock?</em></td>
<td>At twelve o’clock...</td>
<td>sandwich + eat</td>
</tr>
<tr>
<td>11b</td>
<td>Wat voor schoenen dragen mannen het liefst op hun werk?</td>
<td>Op het werk...</td>
<td>sportschoenen</td>
</tr>
<tr>
<td></td>
<td><em>What kind of shoes do men prefer to wear at work?</em></td>
<td>At work...</td>
<td>sneakers</td>
</tr>
<tr>
<td>13a</td>
<td>Heeft Danny te hard gereden?</td>
<td>Ja, en nu...</td>
<td>boete + krijgen</td>
</tr>
<tr>
<td></td>
<td><em>Did Danny drive too fast?</em></td>
<td>Yes, and now...</td>
<td>ticket + get</td>
</tr>
<tr>
<td>13b</td>
<td>Hebben ze al een doelpunt gemaakt?</td>
<td>Nee, dus nu...</td>
<td>Verliezen</td>
</tr>
<tr>
<td></td>
<td><em>Have they scored a goal yet?</em></td>
<td>No, so now...</td>
<td>lose</td>
</tr>
</tbody>
</table>
Appendix C:
GREET treatment target Items (Chapter 3)

<table>
<thead>
<tr>
<th>Sentence ID</th>
<th>Required Answer</th>
</tr>
</thead>
</table>
| 22 (s.1)    | eerst gaat Melvin naar de gemeente  
first Melvin goes to the municipality  |
| 25 (s.1)    | als eerste tillen ze de bank uit de auto  
firstly they lift the sofa out of the car  |
| 30 (s.1)    | in haar handen heeft Karin een lang stuk touw  
in her hands Karin has a long length of rope  |
| 32 (s.1)    | het langst kijkt Melvin naar Marion  
Melvin looks at Marion the longest  |
| 34 (s.1)    | met zijn vieren trekken ze de bank naar boven  
with four they pull the sofa up  |
| 35 (s.1)    | nee want nu kan niemand door de gang lopen  
No, because now nobody can walk through the hallway  |
| 38 (s.1)    | eerst haalt Melvin de doos uit de gang  
first Melvin moves the box out of the hallway  |
| 40 (s.1)    | met drie mensen kunnen ze de bank omhoog trekken  
with three persons they can pull the sofa up  |
| 41 (s.1)    | op de verhuisdoos van Melvin staat zijn naam  
On the box it says Melvin’s name  |
| 44 (s.1)    | eerst gaat Melvin de buurvrouw helpen  
first Melvin helps the neighbor  |
| 46 (s.1)    | in de winkel is Tom een klant aan het helpen  
in the store, Tom is helping a customer  |
| 47 (s.1)    | als Karin weggaat trekken Mo en Marion aan het touw  
when Karin leaves, Mo and Marion are pulling the rope  |
| 48 (s.1)    | bijna vergeet de mevrouw om melk te kopen  
almost the lady forgets to buy milk  |
| 50 (s.1)    | misschien wil de man vervelend doen  
maybe the man wants to be annoying  |
| 57 (s.1)    | aan tafel drinken Mo en Melvin samen thee  
at the table Mo and Melvin drink tea together  |
| 60 (s.1)    | in de supermarkt kan hij suiker kopen  
in the supermarket he can buy sugar  |
| 67 (s.1)    | misschien ligt de brief naast de televisie  
maybe the letter is next to the television  |
| 70 (s.1)    | op zondag om 11 uur gaan ze samen koffie drinken  
on Sunday at 11 o’clock they will have coffee together  |
<table>
<thead>
<tr>
<th>Number</th>
<th>Translation</th>
</tr>
</thead>
</table>
| 74 (s.1) | met een briefje van vijftig betaalt Melvin  
*Melvin pays with a note of fifty* |
| 81 (s.2) | voor zes uur moet de brief op de post  
*before six o'clock the letter must be posted* |
| 83 (s.2) | morgen kan anne de brief meenemen  
*tomorrow Anne can take the letter with her* |
| 89 (s.2) | onder het bed zoeken Tom en Karin naar de brief  
*Tom and Karin are looking under the bed for the letter* |
| 95 (s.2) | nee dus vraagt ze dat aan Karin  
*No, so she asks Karin* |
| 99 (s.2) | nee gisteren heeft Mo slecht geslapen  
*No, yesterday Mo slept poorly* |
| 102 (s.2) | daar staat dat lamsvlees goedkoop is  
*there it says that lamb is cheap* |
| 105 (s.2) | op een kaart wijst Mo de supermarkt aan  
*on a map Mo shows where the supermarket is* |
| 113 (s.2) | in zijn winkel verkoopt Tom boter en kaas  
*in his store Tom sells butter and cheese* |
| 116 (s.2) | op de klok aan de muur is het half tien  
*on the clock on the wall it is nine thirty* |
| 117 (s.2) | over een uur heeft Tom lamsvlees  
*in an hour Tom will have lamb* |
| 118 (s.2) | als Tom weggaat vraagt hij aan Karin om te helpen  
*when Tom leaves he asks Karin to help* |
| 119 (s.2) | met de fiets gaat Tom naar de supermarkt  
*Tom is going to the supermarket by bicycle* |
| 123 (s.2) | voor Melvin haalt Tom de boodschappen  
*Tom is getting Melvin's groceries* |
| 126 (s.2) | voor Tom was een mevrouw aan de beurt  
*before Tom a lady was being served* |
| 132 (s.2) | bij de slager is lamsvlees duur  
*at the butcher's, lamb is expensive* |
| 135 (s.2) | vanavond maakt Melvin lamsvlees met paprika  
*tonight Melvin is making lamb with paprika* |
| 136 (s.2) | straks wil Melvin ook paprika's kopen  
*later Melvin will also want to buy paprikas* |
| 141 (s.2) | morgen komen de nieuwe paprika's  
*tomorrow the new paprikas will arrive* |
| 142 (s.2) | pas over een maand kun je paprika's kopen  
*Only in a month will you be able to buy paprikas* |
| 143 (s.2) | ja daarom wijst Karin de weg  
*yes, that is why Karin shows him the way* |
| 144 (s.2) | vanavond eet Melvin rijst met lamsvlees en paprika  
*tonight Melvin eats rice with lamb and paprika* |
**Appendix D: Post-test questionnaire (Chapter 3)**

<table>
<thead>
<tr>
<th>No.</th>
<th>Question (response: Likert scale: strongly disagree(1) - strongly agree (5))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ik vond het leuk om met het programma te werken (I enjoyed working with the program)</td>
</tr>
<tr>
<td>2</td>
<td>Oefenen met GREET is een goede manier om mijn Nederlands te verbeteren (Practicing with GREET is a good way to improve my Dutch)</td>
</tr>
<tr>
<td>3</td>
<td>Mijn Nederlands is beter geworden door te oefenen met GREET (My Dutch has improved through practice with GREET)</td>
</tr>
<tr>
<td>4</td>
<td>Het programma gebruiken was moeilijk (Using the program was difficult)</td>
</tr>
<tr>
<td>5</td>
<td>Ik oefen liever met gewone huiswerk oefeningen dan met dit programma (I prefer traditional homework exercises over practice with this program)</td>
</tr>
<tr>
<td>6</td>
<td>Ik vind het vervelend om in een microfoon te spreken (I dislike speaking into a microphone)</td>
</tr>
<tr>
<td>7</td>
<td>De instructies waren duidelijk (The instructions were clear)</td>
</tr>
<tr>
<td>8</td>
<td>De video’s zijn saai (The videos are boring)</td>
</tr>
<tr>
<td>9</td>
<td>De gesprekken in de video’s zijn makkelijk te begrijpen (The conversations in the videos were easy to understand)</td>
</tr>
<tr>
<td>10</td>
<td>De oefening was te lang (The practice session was too long)</td>
</tr>
<tr>
<td>11</td>
<td>Ik heb de video’s al eerder gezien (I have seen the videos before)</td>
</tr>
<tr>
<td>12</td>
<td>De vragen waren saai (The questions were boring)</td>
</tr>
<tr>
<td>13</td>
<td>De vragen waren moeilijk (The questions were difficult)</td>
</tr>
<tr>
<td>15-CF</td>
<td>De feedback van GREET was prettig (The feedback by GREET was nice)</td>
</tr>
<tr>
<td>16-CF</td>
<td>Ik vind het nuttig om feedback te krijgen als ik Nederlands spreek (I think it is useful to receive feedback when I speak Dutch)</td>
</tr>
<tr>
<td>17-CF</td>
<td>Soms zei GREET dat ik een fout maakte als dat niet zo was (Sometimes GREET said that I made an error even though I did not)</td>
</tr>
<tr>
<td>18-CF</td>
<td>Soms maakte ik een fout, maar zei GREET dat het goed was (Sometimes I made an error, but GREET said it was correct)</td>
</tr>
<tr>
<td>15-NOCF</td>
<td>Ik wil dat GREET het zegt als ik een fout maak (I want GREET to tell me when I make an error)</td>
</tr>
<tr>
<td>16-NOCF</td>
<td>Ik merk het zelf als ik fouten maak in mijn Nederlands (I notice it myself when I make mistakes in Dutch)</td>
</tr>
</tbody>
</table>
Appendix E:
Grammaticality judgment task items (Chapter 4 & 5)

<table>
<thead>
<tr>
<th>Pair</th>
<th>V</th>
<th>Stimulus sentence</th>
<th>G/U</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGETS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>In de winkel ziet Laura haar vriendin</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>Van zijn nieuwe leraar Kevin mocht niet bellen</td>
<td>U</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>Na de ruzie hebben Sofie en Anouk het goed gemaakt</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>Om vijf voor negen de bus vertrekt naar Amsterdam</td>
<td>U</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>Als de regen voorbij is, dan fietst Max naar school</td>
<td>G</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>De jongen zit thuis en daar hij wacht hij op Stefanie</td>
<td>U</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>Als Koen naar huis gaat, dan neemt hij bus veertien</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>Aan zijn hond Mark geeft alleen het lekkerste eten</td>
<td>U</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>Voor de buurman maakt Tom een taart</td>
<td>G</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>Voor het recept wij hebben twee tomaten en een ui nodig</td>
<td>U</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>Vanavond komt Thomas bij ons eten</td>
<td>G</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>Door de file Tim was te laat</td>
<td>U</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>Op de markt hebben de meisjes nieuwe schoenen gekocht</td>
<td>G</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>Voordat Lucas mag spelen, hij moet eerst opruimen</td>
<td>U</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>Sanne heeft gevoetbald en nu heeft zij dorst</td>
<td>G</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>Als Rick een nieuwe auto koopt, dan hij kijkt niet naar de prijs</td>
<td>U</td>
</tr>
<tr>
<td>17</td>
<td>A</td>
<td>Anne houdt van lezen en vandaag leest zij een nieuw boek</td>
<td>G</td>
</tr>
<tr>
<td>18</td>
<td>A</td>
<td>Nick is jarig en daarom wij gaan op bezoek</td>
<td>U</td>
</tr>
<tr>
<td>19</td>
<td>A</td>
<td>Jeroen maakt eerst zijn huiswerk en dan drinkt hij thee</td>
<td>G</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>In de grote tuin van Emma veel mooie bomen staan</td>
<td>U</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>In de winkel Laura ziet haar vriendin</td>
<td>U</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>Van zijn nieuwe leraar mocht Kevin niet bellen</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>Na de ruzie Sofie en Anouk hebben het goed gemaakt</td>
<td>U</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Om vijf voor negen vertrekt de bus naar Amsterdam</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>Als de regen voorbij is, dan Max fietst naar school</td>
<td>U</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>De jongen zit thuis en daar wacht hij op Stefanie</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>Als Koen naar huis gaat, dan hij neemt bus veertien</td>
<td>U</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>Aan zijn hond geeft Mark alleen het lekkerste eten</td>
<td>G</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>Voor de buurman Tom maakt een taart</td>
<td>U</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>Voor het recept hebben wij twee tomaten en een ui nodig</td>
<td>G</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>Vanavond Thomas komt bij ons eten</td>
<td>U</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>Door de file was Tim te laat</td>
<td>G</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>Op de markt de meisjes hebben nieuwe schoenen gekocht</td>
<td>U</td>
</tr>
</tbody>
</table>
### The Role of Automatic CF in Learning L2 Grammar

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>B</td>
<td>Voordat Lucas mag spelen, moet hij eerst opruimen</td>
<td>G</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>Sanne heeft gevoetbald en nu zij heeft dorst</td>
<td>U</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>Als Rick een nieuwe auto koopt, dan kijkt hij niet naar de prijs</td>
<td>G</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>Anne houdt van lezen en vandaag zij leest een nieuw boek</td>
<td>U</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>Nick is jarig en daarom gaan wij op bezoek</td>
<td>G</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
<td>Jeroen maakt eerst zijn huiswerk en dan hij drinkt thee</td>
<td>U</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>In de grote tuin van Emma staan veel mooie bomen</td>
<td>G</td>
</tr>
</tbody>
</table>

### FILLERS

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>De student las vier boeken voor de les</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>Moeder pakt drie bord van de tafel</td>
<td>U</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>In de stad mag je niet zo hard rijden</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>Jessica heeft een kaart schrijven</td>
<td>U</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>Wij kunnen deze oefening op de computer doen</td>
<td>G</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>Martijn wil kopen een nieuwe fiets</td>
<td>U</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>Het meisje is acht jaar oud</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>De hondje is niet gevaarlijk</td>
<td>U</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>Jeroen was gisteren in Amsterdam bij zijn ouders</td>
<td>G</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>De zwemmer heeft vorig jaar de eerste prijs</td>
<td>U</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>De kat slaapt overdag veel</td>
<td>G</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>De computers wordt gemaakt</td>
<td>U</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>Naomi eet een sinaasappel als ontbijt</td>
<td>G</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>Mike en zijn vriend de auto maken</td>
<td>U</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>Erik geeft de student een appel en een boterham</td>
<td>G</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>Die meneer rijdt zonder kaartje in tram</td>
<td>U</td>
</tr>
<tr>
<td>17</td>
<td>A</td>
<td>Het is vandaag lekker weer</td>
<td>G</td>
</tr>
<tr>
<td>18</td>
<td>A</td>
<td>Esther heeft een mooi tas</td>
<td>U</td>
</tr>
<tr>
<td>19</td>
<td>A</td>
<td>Mijn broer gaf het boek aan zijn vriend</td>
<td>G</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>Patrick gaf een kaartje voor het concert mij</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>De student las vier boeken voor de les</td>
<td>U</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>Moeder pakt drie borden van de tafel</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>In de stad mag je niet zo hard gereden</td>
<td>U</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Jessica heeft een kaart geschreven</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>Wij kunnen doen deze oefening op de computer</td>
<td>U</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>Martijn wil een nieuwe fiets kopen</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>Het meisje is acht jaar oud</td>
<td>U</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>De hondje is niet gevaarlijk</td>
<td>G</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>Jeroen is gisteren in Amsterdam bij zijn ouders</td>
<td>U</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>De zwemmer had vorig jaar de eerste prijs</td>
<td>G</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>De kat slapen overdag veel</td>
<td>U</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>De computers worden gemaakt</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>Naomi een sinaasappel als ontbijt eet</td>
<td>U</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>Mike en zijn vriend maken de auto</td>
<td>G</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>Erik geeft student een appel en een boterham</td>
<td>U</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>Die meneer rijdt zonder kaartje in de tram</td>
<td>G</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>Het is vandaag lekkere weer</td>
<td>U</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>Esther heeft een mooie tas</td>
<td>G</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
<td>Mijn broer gaf het boek zijn vriend</td>
<td>U</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>Patrick gaf mij een kaartje voor het concert</td>
<td>G</td>
</tr>
</tbody>
</table>

V = version, g/u = grammatical or ungrammatical
Appendix F: Discourse completion task items (Chapter 4 & 5)

<table>
<thead>
<tr>
<th>V</th>
<th>Question</th>
<th>Sentence to complete</th>
<th>Clue words</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGETS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 A</td>
<td>In de boom zit een mooie vogel.</td>
<td>Met zijn camera...</td>
<td>foto + maken</td>
</tr>
<tr>
<td>2 A</td>
<td>Wanneer gaat Patricia op vakantie?</td>
<td>Volgende...</td>
<td>week + Amerika</td>
</tr>
<tr>
<td>3 A</td>
<td>Hoe laat gaat Linda slapen?</td>
<td>Om tien uur 's avonds...</td>
<td>naar bed</td>
</tr>
<tr>
<td>4 A</td>
<td>Wat doet Ans op zaterdag?</td>
<td>Op zaterdag...</td>
<td>de stad + gaan</td>
</tr>
<tr>
<td>5 A</td>
<td>Wil je vandaag werken?</td>
<td>Nee, vandaag...</td>
<td>thuis + blijven</td>
</tr>
<tr>
<td>6 A</td>
<td>Heeft Danny te hard gereden?</td>
<td>Ja, en nu...</td>
<td>boete + krijgen</td>
</tr>
<tr>
<td>7 A</td>
<td>Esther en Danielle kijken een film.</td>
<td>In de film...</td>
<td>goede acteurs + spelen</td>
</tr>
<tr>
<td>8 A</td>
<td>Wat voor mannen werken hier?</td>
<td>Bij de bank...</td>
<td>mannen + rijk</td>
</tr>
<tr>
<td>9 A</td>
<td>Moeder heeft een taart gebakken.</td>
<td>Op de taart ...</td>
<td>15 kaarsen</td>
</tr>
<tr>
<td>10 A</td>
<td>Dit is een mooie vijver.</td>
<td>In de vijver...</td>
<td>eenden + zwemmen</td>
</tr>
<tr>
<td>11 A</td>
<td>Sander is aan de tafel aan het spelen.</td>
<td>Op de tafel ...</td>
<td>auto's + spelen</td>
</tr>
<tr>
<td>12 A</td>
<td>Wat doet Rob morgen op de markt?</td>
<td>Morgen op de markt ...</td>
<td>bloemen + kopen</td>
</tr>
<tr>
<td>13 A</td>
<td>De kinderen hebben vandaag vrij.</td>
<td>maar maandag...</td>
<td>naar school</td>
</tr>
<tr>
<td>14 A</td>
<td>Dennis ziet niks leuks op tv.</td>
<td>Dennis zet de tv uit en dan...</td>
<td>hardlopen</td>
</tr>
<tr>
<td>15 A</td>
<td>Richard is bang voor spinnen.</td>
<td>Als hij een spin ziet, ...</td>
<td>buiten + rennen</td>
</tr>
<tr>
<td>16 A</td>
<td>Het is acht oktober.</td>
<td>Marloes is jarig en nu...</td>
<td>feest + geven</td>
</tr>
<tr>
<td>1 B</td>
<td>Het is buiten heel koud.</td>
<td>Met zijn wame jas ...</td>
<td>buiten + gaan</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>De les begint om 08.30. Hoelang duurt het tot de les begint?</td>
<td>Over...</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>Wat eet Kim om twaalf uur?</td>
<td>Om twaalf uur...</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Jeroen moet naar Londen. Wanneer gaat hij?</td>
<td>Morgen...</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>Ga je mee naar de film?</td>
<td>Nee, vanavond...</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>Wendy moet deze week naar Den Haag.</td>
<td>Dinsdag om elf uur...</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>Bianca houdt van muziek maken.</td>
<td>Op de gitaar...</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>Het regent.</td>
<td>Onder de paraplu...</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>Wat voor schoenen draagt Erik op zijn werk?</td>
<td>Op het werk...</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>Arnhem is een drukke stad.</td>
<td>In Arnhem...</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>Patrick gaat naar zijn vriend in Amsterdam.</td>
<td>In de stad...</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>Wat doen jullie morgen op school?</td>
<td>Morgen op school...</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>Vandaag rijdt Bas in een langzame auto, maar zaterdag...</td>
<td>snelle auto</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>Ellen wil brood kopen maar de bakker is dicht. Wat doet zij?</td>
<td>Ellen loopt weg en dan...</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>Laura is erg laat voor de trein. Als zij heel hard rent, ...</td>
<td>trein + halen</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>Ze hebben geen doelpunten gemaakt.</td>
<td>Paul heeft verloren en nu...</td>
</tr>
</tbody>
</table>

**FILLERS**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>Wat doet deze jongen?</td>
<td>Maarten...</td>
<td>boek</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>Ik vind deze winkel niet duur.</td>
<td>Die trui...</td>
<td>20 euro</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>Koop je die trui?</td>
<td>Ja, hij...</td>
<td>lekker warm</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>Is die mooie auto van jou?</td>
<td>Nee, die...</td>
<td>van de buurman</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>Chantal zit op Nederlandse les.</td>
<td>De leraar...</td>
<td>niet aardig</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>Dit is een oude foto.</td>
<td>Deze vier kinderen...</td>
<td>heel oud</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>De auto is kapot. Kan Frank hem maken?</td>
<td>Ja, Frank kan...</td>
<td>auto</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>Wat gaat Marieke doen?</td>
<td>Marieke gaat...</td>
<td>kopen</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>Wat wil Suzan doen?</td>
<td>Suzan wil... kado + geven</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>Het station is dichtbij, mevrouw.</td>
<td>Het is... vijf minuten + lopen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>Joyce is hard aan het werken.</td>
<td>Zij wil... vakantie + nemen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>Je mag hier met de auto 50 rijden,</td>
<td>maar... veel harder</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>13</strong></td>
<td>Peter wil op de markt sinaasappels kopen.</td>
<td>Hij wil dat omdat hij... sinaasappels + lekker</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>14</strong></td>
<td>Het is al negen uur. Waar blijft Paul?</td>
<td>Ik denk dat... te laat</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>15</strong></td>
<td>Waarom zijn de winkels dicht?</td>
<td>De winkels zijn dicht omdat... zondag</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>16</strong></td>
<td>Dat zijn leuke schoenen.</td>
<td>Ik denk dat... kopen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1</strong></td>
<td>Tim is ziek. Wanneer komt de dokter?</td>
<td>De dokter... Vanmiddag</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Ga je uit vanavond?</td>
<td>Nee, ik... thuis</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Zijn dit alle studenten?</td>
<td>Ja, deze klas... klein</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>De benzine is bijna op. Kunnen we wel tanken?</td>
<td>Ja, veel benzinepompen... laat + open</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Hans kan niet met de fiets.</td>
<td>Zijn band... lek</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Wat zijn deze mannen?</td>
<td>Deze mannen... verhuizers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Gaat Sandra vandaag iets kopen?</td>
<td>Ja, Sandra gaat... een nieuwe broek</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>Wat moet Jan morgen doen?</td>
<td>Jan moet... wassen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>Wat wil Karin doen?</td>
<td>Karin wil... bromfiets + kopen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>Zie je de kinderen?</td>
<td>Ja, ze zijn... buiten + spelen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>Frank moet naar school, maar het is te ver om te lopen.</td>
<td>Hij wil... bus + nemen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>Ik wil die schoenen kopen, maar...</td>
<td>maar... heel duur</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>13</strong></td>
<td>Is het buiten mooi weer?</td>
<td>Het is koud, maar... zon + schijnen</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>14</strong></td>
<td>Waar is Nicole?</td>
<td>Ik denk dat... ziek</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>15</strong></td>
<td>Gaat Mark de jas ruilen?</td>
<td>Ja, Mark wil hem ruilen omdat... klein</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>16</strong></td>
<td>De docent vraagt iets aan de klas.</td>
<td>Niels denkt dat... Het antwoord + weten</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V= version. Each test item is accompanied by a picture.
### Appendix G: Pre-test questionnaire

<table>
<thead>
<tr>
<th>Question</th>
<th>Type of response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Leeftijd:</td>
<td>open</td>
</tr>
<tr>
<td>2  Geslacht:</td>
<td>man/vrouw</td>
</tr>
<tr>
<td>3  Waar leer je Nederlands?</td>
<td>closed list of participating institutes</td>
</tr>
<tr>
<td>4  CEF Niveau van Nederlands:</td>
<td>open</td>
</tr>
<tr>
<td>5  Niveau van opleiding:</td>
<td>hoger onderwijs (bv. hogeschool, universiteit)</td>
</tr>
<tr>
<td>6  Land van herkomst</td>
<td>open</td>
</tr>
<tr>
<td>7  Moedertaal</td>
<td>open</td>
</tr>
<tr>
<td>8  Spreek je meer talen dan Nederlands en je moedertaal? Zo ja, welke?</td>
<td>open</td>
</tr>
<tr>
<td>9  Ik woon al .......... jaar in Nederland</td>
<td>open</td>
</tr>
<tr>
<td>10 Waarom ben je naar Nederland gekomen?</td>
<td>open</td>
</tr>
<tr>
<td>11 Ik leer al .......... jaar Nederlands.</td>
<td>open</td>
</tr>
<tr>
<td>12 Ik gebruik de computer:</td>
<td>elke dag</td>
</tr>
<tr>
<td>13 Geef aan hoeveel tijd je ongeveer per dag bezig bent met Nederlands lezen en schrijven:</td>
<td>0 tot 30 min</td>
</tr>
<tr>
<td>14 Geef aan hoeveel tijd je ongeveer per dag bezig bent met Nederlands luisteren en spreken:</td>
<td>0 tot 30 min</td>
</tr>
<tr>
<td>15 Ik verleen mijn toestemming om mijn gegevens en de opgenomen data voor onderzoek te gebruiken. I give my consent that my information and the recorded data may be used for research purposes.</td>
<td>Ja/Nee</td>
</tr>
</tbody>
</table>
Appendix H:
Post-test questionnaire items (Chapter 4 & 5)

<table>
<thead>
<tr>
<th>Shared questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Ik vond het ______ om met GREET te werken (Saai-Leuk)</td>
</tr>
<tr>
<td>I thought it was ______ to work with GREET (boring - fun)</td>
</tr>
<tr>
<td>1. Oefenen met GREET is ______ om mijn Nederlands te verbeteren (Slecht-Goed)</td>
</tr>
<tr>
<td>Practicing with GREET is a (bad – good) way to improve my Dutch</td>
</tr>
<tr>
<td>2. Mijn Nederlands is ______ geworden door te oefenen met GREET (Slechter-Beter)</td>
</tr>
<tr>
<td>My Dutch has ______ as a result of practicing with GREET (worsened - improved)</td>
</tr>
<tr>
<td>3. Het programma gebruiken was ______ (Moeilijk-Makkelijk)</td>
</tr>
<tr>
<td>Using the program was ______ (difficult - easy)</td>
</tr>
<tr>
<td>4. De instructies waren ______ (Onduidelijk-Duidelijk)</td>
</tr>
<tr>
<td>The instructions were ______ (unclear - clear)</td>
</tr>
<tr>
<td>5. De video's zijn ______ (Saai-Leuk)</td>
</tr>
<tr>
<td>The videos were ______ (boring - fun)</td>
</tr>
<tr>
<td>6. De video's zijn ______ te begrijpen (Moeilijk-Makkelijk)</td>
</tr>
<tr>
<td>The videos are ______ to understand (difficult - easy)</td>
</tr>
<tr>
<td>7. De vragen waren ______ (Saai-Leuk)</td>
</tr>
<tr>
<td>The questions were ______ (boring - fun)</td>
</tr>
<tr>
<td>8. De vragen waren ______ (Moeilijk-Makkelijk)</td>
</tr>
<tr>
<td>The questions were ______ (difficult - easy)</td>
</tr>
<tr>
<td>9. Ik leer het best Nederlands spreken wanneer mijn fouten worden gecorrigeerd</td>
</tr>
<tr>
<td>(Oneens-Mee eens)</td>
</tr>
<tr>
<td>The best way for me to learn to speak Dutch is when my errors are corrected</td>
</tr>
<tr>
<td>(disagree - agree)</td>
</tr>
<tr>
<td>10. Ik vind het ______ wanneer mijn taalfout wordt uitgelegd met een regel</td>
</tr>
<tr>
<td>I think it is ______ when my language error is explained with a rule (not useful-</td>
</tr>
<tr>
<td>useful)</td>
</tr>
</tbody>
</table>

Questions per experiment condition

<table>
<thead>
<tr>
<th>NOCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. De oefening was te ______ (Lang-Kort)</td>
</tr>
<tr>
<td>The exercise was too ______ (long - short)</td>
</tr>
<tr>
<td>12. Als ik antwoord aan GREET, dan vind ik spreken ______ dan typen</td>
</tr>
<tr>
<td>(Slechter-Beter)</td>
</tr>
<tr>
<td>When responding to GREET, I ______ speaking to typing (disprefer - prefer)</td>
</tr>
<tr>
<td>13. Als ik fouten maak in het Nederlands, dan merk ik dat ______</td>
</tr>
<tr>
<td>(Nooit-Altijd)</td>
</tr>
<tr>
<td>If I make mistakes in Dutch, I notice that ______ (Never - always)</td>
</tr>
<tr>
<td>14. Ik wil graag dat mijn fouten meteen worden gecorrigeerd (Oneens -Mee eens)</td>
</tr>
<tr>
<td>I prefer to have my errors corrected immediately (disagree - agree)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>CF</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
</tbody>
</table>
## Appendix J:
Treatment Items (Chapter 4 & 5)

<table>
<thead>
<tr>
<th>clip no.</th>
<th>q.no.</th>
<th>Question - Targeted response</th>
<th>T / F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Session 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Wat gaat Melvin doen? - Melvin gaat verhuizen</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Melvin gaat verhuizen. Wat doet hij eerst? Eerst gaat Melvin naar de gemeente</td>
<td>T</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Melvin rijdt in een busje. Wat is de kleur van het busje? De kleur van het busje is wit.</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>Er loopt een man naar Melvin toe. Wie is dat? Dat is zijn oom Mo.</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>Wat tillen Mo en Melvin als eerste uit de auto? Als eerste tillen ze de bank uit de auto</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>De bank is te zwaar voor Mo en Melvin. Mo vraagt om hulp. Hoe kan hij dat vragen? Kunnen jullie mij met deze bank helpen?</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>Waar moet de bank naartoe? De bank moet naar boven</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>Op welke verdieping gaat Melvin wonen? Melvin gaat op de tweede verdieping wonen</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>Wie heeft Mo om hulp gevraagd? Mo heeft Tom, Karin en Marion om hulp gevraagd</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>Wat heeft Karin in haar handen? In haar handen heeft Karin een lang touw.</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>Waar gaat Melvin wonen? Melvin gaat bij zijn oom Mo wonen</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>Melvin ontmoet drie mensen. Naar wie kijkt Melvin het langst? Het langst kijkt Melvin naar Marion</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>Melvin gaat op de tweede verdieping wonen. Hoe moet de bank naar boven? De bank moet omhoog worden getrokken</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>Met hoeveel mensen trekken ze de bank omhoog? Met zijn vieren trekken ze de bank naar boven</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>Melvin zet een doos in de gang. Is dat een probleem? Ja, nu kan niemand door de gang lopen</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>De doos staat in de weg. Een vrouw met een kind kan er niet langs. Wie is dat? Dat is de buurvrouw en zij heet Alice Timmerman</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>Wat vraagt de buurvrouw aan Melvin?</td>
<td>F</td>
</tr>
</tbody>
</table>
Is die doos van jou?

5  18 Melvin moet helpen op het balkon. Maar wat doet hij?
Eerst haalt Melvin de doos uit de gang T

5  19 Waarom laat Tom het touw los?
Omdat hij een vrouw in de winkel moet helpen F

5  20 Ze trekken nu met drie mensen aan het touw. Kan dat?
Ja, met drie mensen kunnen zij de bank omhoog trekken T

5  21 Wat staat er op de verhuisdoos van Melvin?
Op de verhuisdoos van Melvin staat zijn naam T

6  22 Hoeveel moet de mevrouw betalen?
Zij moet negen dertig betalen F

6  23 Hoeveel geld krijgt zij terug?
Zij krijgt zeventig cent terug F

6  24 Melvin moet gaan helpen met de bank. De buurvrouw is iets
vergeten. Wat doet Melvin?
Eerst gaat Melvin de buurvrouw helpen T

7  25 Drie mensen trekken aan het touw: Karin, Marion en Mo.
Waarom laat Karin het touw los?
Omdat Tom roept dat zij de telefoon moet opnemen F

7  26 Waarom kan Tom de telefoon niet opnemen?
In de winkel moet Tom eerst een klant helpen T

7  27 Karin gaat de winkel in. Wie trekken nu aan het touw?
Als Karin weggaat trekken Mo en Marion aan het touw T

7  28 De mevrouw in de winkel vergeet bijna iets. Wat is dat?
Bijna vergeet de mevrouw om melk te kopen T

8  29 Waarom kan Marion Melvin niet horen?
Een man met een brommer maakt veel lawaai F

8  30 Waarom maakt die man zoveel lawaai, denk je?
Misschien wil de man verval dan toen F

8  31 Wat roept Melvin naar Marion en Mo?
Hij roept dat de bank hoger moet F

9  32 Waarom laat Marion het touw los?
Omdat zij Melvin niet kan verstaan F

9  33 Mo heeft het touw nog wel vast. Wat gebeurt er met hem?
Mo wordt omhoog getrokken F

9  34 Waarom kan hij de bank niet omhoog houden?
Dat kan niet omdat de bank heel zwaar is F

9  35 Gelukkig valt Mo niet hard. Waarom niet?
Mo valt niet hard omdat hij op de bank valt F

10 36 In het filmpje zitten Mo en Melvin aan tafel. Waar praten ze
over?
Ze willen de buren uitnodigen voor een kopje koffie T

10 37 Wat drinken Mo en Melvin?
| 10 | 38 | Doet Melvin suiker in zijn thee? 
Nee, dat kan niet want de suiker is op | F |
| 10 | 39 | Wanneer willen ze de buren uitnodigen? 
Ze willen de buren uitnodigen in het weekend | F |
| 11 | 40 | De suiker is op. Waar kan Melvin nieuwe suiker kopen? 
In de supermarkt kan hij suiker kopen | T |
| 11 | 41 | Mo vraagt Melvin om nog iets te kopen. Wat is dat? 
Mo vraagt aan Melvin om melk te kopen | F |
| 11 | 42 | Naar welke supermarkt gaat hij? 
Melvin gaat naar de supermarkt van Tom en Karin | F |
| 11 | 43 | Wat zoekt Tom? 
Hij zoekt een brief van de school van Anne | F |
| 11 | 44 | Wie is Anne? 
Anne is de dochter van Tom en Karin | F |
| 12 | 45 | Tom zoekt een brief. Waar ligt de brief volgens Karin? 
Volgens Karin ligt de brief in de woonkamer | T |
| 12 | 46 | Waar kan de brief nog meer liggen volgens Karin? 
De brief kan ook in de keuken liggen | F |
| 12 | 47 | En waar kan de brief nog meer liggen? 
Misschien ligt de brief naast de televisie | T |
| 12 | 48 | Wat vraagt Melvin aan Karin? 
Melvin vraagt of zij op visite komen | F |
| 12 | 49 | Karin kan zondag. Kan Tom ook op zondag komen? 
180 dat weet 180 Karin niet | F |
| 12 | 50 | Wanneer gaan ze samen koffie drinken? 
Op zondag om 11 uur gaan zij samen koffie drinken | T |
| 14 | 51 | Melvin koopt koekjes en melk. Wat koopt hij nog meer? 
Hij koopt ook koffie. | F |
| 14 | 52 | Melvin is iets vergeten. Hij zou twee dingen doen, welke twee? 
Melvin wilde suiker kopen en de buren uitnodigen. | F |
| 14 | 53 | Wat is Melvin vergeten? 
Melvin is vergeten suiker te kopen | F |
| 14 | 54 | De boodschappen kosten zes gulden vijftigentwintig. Hoe betaalt Melvin? 
Met een briefje van vijftig betaalt hij de boodschappen | T |
| 14 | 55 | Wat is een gulden? 
Vroeger moest je in Nederland met guldens betalen | T |
| 14 | 56 | Hoe betaal je nu? 
Nu betaal je in Nederland met Euro's | T |
| 15 | 57 | Als Melvin thuiskomt hangt hij zijn jas op. Welke kleur 180anks zijn jas? 
De kleur van zijn jas is grijs | F |
Appendix | 181

15 58  Tom en Karin hebben ruzie, zegt Melvin. Waarom hebben ze ruzie?
      Tom is een belangrijke brief kwijt  F
15 59  Hebben Tom en Karin ook echt ruzie?
      Tom wil graag de brief vinden en Karin snapt niet waarom  F
15 60  Mo vraagt of de buren op visite komen. Wat zegt Melvin?
      Karin komt wel, maar misschien wil Tom niet  F
16 61  Tom is een brief aan het zoeken. Waar zoekt hij?
      Tom ligt op de grond en zoekt onder het bed.  F
16 62  Waar schrikt Karin van?
      Van Tom, omdat hij haar been pakt  F
16 63  Waarom willen zij de brief snel vinden?
      Als ze de brief snel vinden, 181ank an Tom hem op tijd versturen  F
16 64  Hoe laat moet hij op de post?
      Voor zes uur moet de brief op de post  T
16 65  Hoe laat is het?
      Het is al kwart over vijf  F

Session 2

17 66  Melvin en Mo praten over Tom. Waarom denkt Melvin dat het niet goed gaat met Tom?
      Tom doet vreemd en lijkt nerveus  F
17 67  Waardoor komt dat, denkt Mo?
      Mo denkt dat Tom misschien te hard werkt  F
17 68  Hoe komt Tom aan de winkel?
      De winkel was van de vader van Tom  F
17 69  Mo kijkt op het horloge van Melvin. Hoe laat is het?
      Het is bijna half zes  F
17 70  Mo moet weg omdat het bijna half zes is. Wat doet Melvin?
      Melvin blijft lekker zitten  F
18 71  Tom en Karin zoeken onder het bed. Wat zoeken ze daar?
      Onder het bed zoeken Tom en Karin naar de brief  T
18 72  Wie is het meisje dat binnenkomt?
      Dat is Anne, zij is de dochter van Tom en Karin  F
18 73  Wat staat er in de brief?
      Tom en Karin moeten naar de school van Anne komen  F
18 74  Hoe lang duurt dat gesprek?
      Het gesprek duurt ongeveer tien minuten.  F
19 75  Wat is Tom aan het doen?
      Tom zoekt in zijn agenda naar een vrije avond  F
19 76  Wie is er nerveus voor dat gesprek?
      Tom is nerveus voor het gesprek  F
20 77  Waarom zit Melvin in het begin naast de koelkast?  F
<table>
<thead>
<tr>
<th>Page</th>
<th>Line</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>78</td>
<td>Melvin maakt een lijst met boodschappen. Waarom maakt hij een lijst? Omdat Melvin boodschappen wil gaan doen.</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>Heeft Mo gisteren niet goed geslapen? Nee, gisteren heeft Mo slecht geslapen.</td>
</tr>
<tr>
<td>21</td>
<td>81</td>
<td>Vindt Mo lamsvlees met paprika lekker? Mo zegt dat hij het lekker vindt.</td>
</tr>
<tr>
<td>21</td>
<td>82</td>
<td>Gaat Mo of Melvin koken? Vanavond gaat Melvin het eten koken.</td>
</tr>
<tr>
<td>21</td>
<td>83</td>
<td>Mo leest de krant. Wat ziet hij daar? In de krant ziet Mo een aanbieding voor lamsvlees.</td>
</tr>
<tr>
<td>21</td>
<td>84</td>
<td>Waarom zegt Mo dat Melvin naar de supermarkt moet gaan? Omdat het lamsvlees daar goedkoper is dan bij de slager.</td>
</tr>
<tr>
<td>22</td>
<td>85</td>
<td>Gaat Melvin boodschappen doen in de winkel van Tom en Karin? Nee, hij wil naar de supermarkt gaan.</td>
</tr>
<tr>
<td>22</td>
<td>86</td>
<td>Hoe legt Mo uit waar die andere supermarkt is? Op een kaart wijst Mo de supermarkt aan.</td>
</tr>
<tr>
<td>22</td>
<td>87</td>
<td>Weet Melvin nu waar de supermarkt is? Nee, hij gaat naar de winkel van Tom en Karin.</td>
</tr>
<tr>
<td>22</td>
<td>88</td>
<td>Wat gaat hij in de winkel van Tom en Karin doen? Melvin gaat vragen hoe hij bij de supermarkt komt.</td>
</tr>
<tr>
<td>23</td>
<td>89</td>
<td>Wat vraagt Melvin aan Tom? Melvin vraagt aan Tom waar de supermarkt is.</td>
</tr>
<tr>
<td>23</td>
<td>90</td>
<td>Hoe vaak moet hij vragen waar de supermarkt is? Melvin moet het twee keer vragen.</td>
</tr>
<tr>
<td>23</td>
<td>91</td>
<td>Waarom is Tom verbaasd dat Melvin dat vraagt? Omdat Melvin in de supermarkt van Tom staat.</td>
</tr>
<tr>
<td>23</td>
<td>92</td>
<td>Tom pakt potten uit een blauwe kist. Wat is dat? In de blauwe kist zitten potten met appelmoes.</td>
</tr>
<tr>
<td>24</td>
<td>94</td>
<td>Verkoopt Tom ook boter of kaas? In zijn winkel verkoopt Tom boter en kaas.</td>
</tr>
<tr>
<td>24</td>
<td>95</td>
<td>Melvin vraagt om lamsvlees. Heeft Tom dat, denk je? Ja, Tom zegt dat hij alles heeft.</td>
</tr>
<tr>
<td>25</td>
<td>97</td>
<td>Hoe laat is het als Melvin de winkel in komt? Op de klok aan de muur is het half tien.</td>
</tr>
<tr>
<td>25</td>
<td>98</td>
<td>Wanneer heeft Tom lamsvlees?</td>
</tr>
</tbody>
</table>
Over een uur heeft Tom lamsvlees

<table>
<thead>
<tr>
<th></th>
<th>25</th>
<th>99</th>
<th>Blijft er iemand in de winkel als Tom weggaat?</th>
<th>Als Tom weggaat vraagt hij aan Karin om te helpen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26</td>
<td>100</td>
<td>Hoe gaat Tom weg?</td>
<td>Met de fiets gaat Tom naar de supermarkt</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>101</td>
<td>Wat is een winkelwagentje?</td>
<td>Dat is een kar voor de boodschappen</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>102</td>
<td>Waarom moet hij geld wisselen?</td>
<td>Tom heeft geen kleingeld voor het winkelwagentje</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>103</td>
<td>Wat doet Tom?</td>
<td>Tom wil vlees gaan kopen</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>104</td>
<td>Waarom wil Tom lamsvlees gaan kopen?</td>
<td>Als hij vlees koopt, dan kan hij het aan Melvin verkopen</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>105</td>
<td>Tom is druk bezig. Wat doet Melvin?</td>
<td>Melvin leest thuis de krant</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>106</td>
<td>Dus voor Melvin is dit wel fijn?</td>
<td>Ja, omdat Tom zijn boodschappen haalt</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>107</td>
<td>Is Tom snel aan de beurt bij de slager?</td>
<td>Voor Tom helpt de slager eerst een mevrouw</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>108</td>
<td>Welk nummer heeft Tom bij de slager?</td>
<td>Tom heeft nummer achttien</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>109</td>
<td>De supermarkt heeft niet genoeg vlees. Wat doet Tom?</td>
<td>Tom rent snel de winkel uit</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>110</td>
<td>Wat zegt de vrouw achter de kassa als Tom voorbij rent?</td>
<td>Tot ziens en een prettige dag verder</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>111</td>
<td>Waar is Tom heen gefietst?</td>
<td>Tom is naar de slager gefietst</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>112</td>
<td>Is lamsvlees duur bij de slager?</td>
<td>Bij de slager is lamsvlees duur</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>113</td>
<td>Karin helpt een vrouw in de winkel. Wat koopt zij?</td>
<td>De vrouw wil twee paprika’s kopen</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>114</td>
<td>Waarom heeft ze geluk, volgens Karin?</td>
<td>Omdat zij de laatste paprika’s koopt</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>115</td>
<td>Wat gaat Melvin vanavond koken?</td>
<td>Vanavond koopt Melvin lamsvlees met paprika</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>116</td>
<td>Weet je wat er zal gebeuren?</td>
<td>Straks wil Melvin ook paprika’s kopen</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>117</td>
<td>Hoeveel betaalt de mevrouw voor de laatste twee paprika’s?</td>
<td>Twee paprika’s kosten samen twee gulden</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>118</td>
<td>Karin moest lang op Tom wachten. Wat zegt ze?</td>
<td>ben je daar eindelijk?</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>119</td>
<td>Karin zegt tegen Tom dat de paprika’s op zijn. Hoort Tom dat?</td>
<td>Nee, Tom hoort dat niet</td>
</tr>
</tbody>
</table>
184 | The Role of Automatic CF in Learning L2 Grammar

| 31 | 120 | Tom is blij als Melvin de winkel binnenkomt. Waarom? Omdat hij het lamsvlees aan Melvin kan verkopen | F |
| 31 | 121 | Karin bestelt nieuwe paprika’s. Wanneer komen die? Morgen komen de nieuwe paprika’s | F |
| 31 | 122 | Tom liegt. Wat zegt Tom tegen Melvin? Pas over een maand kun je paprika’s kopen | T |
| 31 | 123 | Karin helpt Melvin. Waarom zegt ze waar de grote supermarkt is? In de supermarkt verkopen ze paprika’s | T |
| 31 | 124 | Op het einde komt alles goed. Wat eet Melvin vanavond? Vanavond eet Melvin rijst met lamsvlees en paprika | T |

T/F = Target (inversion) or Filler sentence
**Nederlandse samenvatting**

**Inleiding**

Bij het leren van een tweede taal hoort het maken van taalfouten. Een gesprekspartner of een leraar kan die fouten aanduiden door het geven van correctieve feedback (CF). Maar heeft het geven van feedback effect op het leren van een tweede taal? Met name voor het leren van spreekvaardigheid is dit al lang een belangrijk punt van discussie. De rol van correctieve feedback heeft implicaties voor theorieën over tweedetaalverwerving. Bijvoorbeeld, als CF effectief is, dan laat dit zien dat taalverwerving niet een uitsluitend onbewust proces is; of als een bepaald type CF het meest effectief blijkt, dan zeggen de karakteristieken van dit CF-type iets over de relevante processen voor taal leren. Het is echter tot dusver erg moeilijk gebleken de effecten van CF op spreekvaardigheid aan te tonen. Ontwikkelingen op het gebied van spraaktechnologie maken het mogelijk om spreekoefeningen op de computer aan te bieden en deze bieden belangrijke nieuwe mogelijkheden voor onderzoek naar CF.

Voor het onderzoek in dit proefschrift ontwikkelden wij GREET: een computerprogramma met spraakherkenning dat automatische CF kan geven op het gesproken Nederlands van tweedetaalleerders en dat bovendien alle interacties van de leerder met de computer registreert. Met behulp van GREET konden wij gedetailleerde taalexperimenten uitvoeren waarbij de effecten van spreekoefeningen met en zonder CF konden worden vergeleken. In drie gedetailleerde studies met GREET onderzochten wij de effecten van CF op verbetering van spreekvaardigheid, in relatie tot leerer gedrag en individuele verschillen tussen leerders. Dit proefschrift presenteert onderzoek naar de rol van CF in de ontwikkeling van spreekvaardigheid. In de analyses besteden we aandacht aan individuele verschillen tussen leerders en aan het verloop van het leerproces. In hoofdstuk 1 wordt een algemene introductie gegeven. Hoofdstuk 2 bespreekt de noodzaak van gecontroleerd onderzoek voor tweedetaalverwervingsonderzoek en de mogelijkheid dit te doen met een computersysteem dat automatische feedback geeft op spreekvaardigheid. Een dergelijk systeem (GREET) wordt gepresenteerd in hoofdstuk 3 en dit wordt ingezet voor een taalexperiment. Hoofdstuk 4 behandelt een vervolgexperiment waarin meer nadruk wordt gelegd op de verschillen tussen leerders in taalvaardigheid en in leerer gedrag. In hoofdstuk 5 wordt GREET ingezet om het taal leren van leerders met verschillende opleidingsniveaus (hoog- en laagopgeleid) te vergelijken. Hoofdstuk 6 is een uitgebreide discussie van de resultaten in de verschillende experimenten, trekt conclusies en doet aanbevelingen voor CF-onderzoek.

Voordat een samenvatting per hoofdstuk wordt gegeven, volgt hier eerst achtergrond over CF en de relevantie van CF voor tweedetaalverwervingsonderzoek.
Verder zal ook de onderzoeksopzet met GREET nader toegelicht worden, omdat deze in de hoofdstukken 3,4,5 gebruikt wordt.

Correctieve feedback

Volwassen leerders van een tweede taal bereiken vrijwel nooit een taalniveau dat vergelijkbaar is met een moedertaalspreker. Doorgaans blijven zij op verschillende niveaus daaronder, soms ook ondanks volledige onderdompeling in die tweedetaalomgeving. Een oorzaak voor het onvolledig leren van de tweede taal is dat sommige (fonologische, morfologische, grammaticale) elementen uit de tweede taal niet (voldoende) opgemerkt worden door de leerder. Om een element te kunnen leren is het nodig dat het element op een of andere manier wordt opgemerkt en verwerkt. Een element van de tweede taal kan onopgemerkt blijven door interferentie van de moedertaal, door redundantie van de vorm voor de betekenis, of door een ondoorzichtige relatie tussen de vorm en de betekenis. Met CF kan de aandacht van de leerder nadrukkelijk gericht worden op een bepaalde vorm en zo het leren mogelijk maken.

Er zijn verschillende manieren waarop feedback kan worden gegeven en elk van deze vormen heeft voor- en nadelen. Zo kan een fout expliciet worden aangeduid (expliciete feedback) of alleen worden gesuggereerd (impliciete feedback). Daarnaast kan CF de correcte vorm geven en de fout verbeteren (input leveren) of de correcte vorm onthouden en de leerder zelf de fout laten herstellen (op de productie van output gericht). De vorm van de CF is van invloed op de effectiviteit voor taal leren, maar het is nog onduidelijk op welke manier. Kennis van wanneer en hoe CF effectief is, is echter van belang om de verschillende perspectieven in tweedetaalverwervingstheorie te kunnen toetsen. Sommige theorieën stellen dat input-leverende CF het meest effectief is, terwijl andere theorieën juist aannemen dat het van belang is dat leerders de fout zelf corrigeren en/of de juiste vorm produceren.

Onderzoek naar correctieve feedback

Omdat CF zeer relevant is voor taalonderwijs en voor taalverwervingstheorieën is er veel onderzoek naar verricht. Empirisch onderzoek naar de rol van CF voor spreekvaardigheid kan de voorspellingen van de verschillende theorieën van tweedetaalverwerving toetsen. Meta-analyses laten over het algemeen een positief effect voor CF zien, maar er is wel veel variabiliteit tussen onderzoeken. Het effect van CF wordt onder meer beïnvloed door het type oefening, de gesprekspartner, de setting van het onderzoek en het type CF. Daarnaast zijn er verschillen in hoe de onderzoeken hun type CF operationaliseren, waardoor de resulterende effecten moeilijk te vergelijken zijn. Ook zijn er verschillen in hoe de effectiviteit wordt gemeten. Deze bevindingen samen hebben geresulteerd in een terugkerende vraag in de literatuur om onderzoeken zo op te zetten dat ze op elkaar kunnen voortbouwen, liefst ook door replicatiestudies. Alleen door experimenten die duidelijk zijn
Samenvatting | 187

gedefinieerd en waar (experimentele) variabelen gecontroleerd worden is dit mogelijk. Een variabele die niet kan worden gecontroleerd, en waarmee dus rekening moet worden gehouden, zijn de individuele verschillen tussen leerders. Als experimenten gecontroleerd zijn, kunnen individuele verschillen systematisch vergeleken worden.

**Invloed van correctieve feedback op het leerproces**

Het gedrag van leerders tijdens een oefening kan veel extra informatie opleveren voor de verklaring van pre-posttoetsscores. CF is de reactie op een taalfout, die weer een effect kan hebben op het taalgedrag van de leerder. Dit effect kan van korte duur zijn (bijvoorbeeld reparatie van de fout) of van langere duur, bijvoorbeeld in de vorm van een afnemende frequentie van die fout in het taalgebruik van de leerder. Het gedrag van de leerder kan derhalve inzicht geven in het effect van CF. Verschillende processen kunnen het gevolg zijn van CF: de fout opmerken, de fout corrigeren, de correcte vorm produceren en het verwerken van taalinput. Door het gedrag van leerders te monitoren kan hierover informatie worden vergaard. Deze informatie kan worden gebruikt om te begrijpen hoe leerdergedrag gerelateerd is aan leeruitkomsten.

**Individuele verschillen en correctieve feedback**

Tot dusver is er in onderzoek nog niet veel aandacht besteed aan individuele kenmerken en de relatie met een CF-effect. Onderzoeken hebben zich vaak gericht op gemiddelde leeruitkomsten van groepen. Sommige studies vonden relaties van CF met individuele kenmerken zoals taalvaardigheidsniveau, werkgeheugen, motivatie en spanning. Het aanpassen van het type CF aan de leerder kan waarschijnlijk de effectiviteit van de CF verhogen. De behoeften van de leerder kunnen evenwel veranderen, bijvoorbeeld doordat de taalvaardigheid verbetert of de motivatie afneemt. Daarom is het noodzakelijk om naast leeruitkomsten ook te kijken naar het leergedrag: wat doet een leerder als reactie op CF? De combinatie van deze informatie geeft een meer volledig beeld van de effectiviteit van CF. En daarmee kan de rol van CF worden onderzocht in interactie met individuele verschillen. In hoofdstuk 4 en 5 wordt ingegaan op de individuele verschillen taalvaardigheid en opleidingsniveau.

**GREET**

Met GREET oefenden leerders Nederlands spreken met en zonder CF. Hoofdstuk 3 gaat in detail in op het ontwerp en de evaluatie van GREET. Alle experimenten in dit proefschrift werden afgenomen volgens dezelfde procedure. De leerders logden in op onze website waar zij in sequentie hun taken voor die dag kregen aangeboden. Deze bestonden uit een korte vragenlijst, twee taalvaardigheidstoetens en 45
minuten oefenen met spreken. In de tweede experimentssessie was de volgorde omgekeerd.

Automatische spraakherkenning is moeilijker naarmate de spreker de taal minder beheerst. Om betrouwbare feedback te geven op taal.leiders van verschillende taalachtergronden moet bij het ontwerp van de oefening rekening worden gehouden met de beperkingen van automatische spraakherkenning. Om methodologische en didactische redenen zijn de eisen voor de precisie van de CF hoog. Hierdoor was het noodzakelijk het aantal mogelijke antwoorden van de leerder te beperken. De oefeningen in GREET gaven daarom leerders woorden of delen van zinnen om hun antwoord mee te construeren. In deze vorm kon geoefend worden met Nederlandse woordvolgorde en onze grammaticale target, waarop de spraakherkennings betrouwbare CF kon geven. In de oefening keken de leerders naar een kort (30 seconden) filmfragment en beantwoordden daarna vragen over het fragment op beginnersniveau (A2-niveau in het Europese referentiekader (CEF)).

Als de leerder (in de CF-conditie) een fout maakt, toont GREET een rood scherm met daarin de melding “Dat klopt niet. Probeer het nog eens”. De leerder moet dan de vraag opnieuw beantwoorden. Bij de tweede keer fout geeft GREET een hint. Deze hint houdt in dat het eerstvolgende woordblok op de juiste plek in de zin is gezet. Na vier keer foutief proberen krijgt de leerder de mogelijkheid om de vraag over te slaan.

Het meten van veranderingen in taalvaardigheid als gevolg van de spreekoeefening is een belangrijk onderdeel van dit onderzoek. Een argument tegen het nut van CF voor taal leren is dat deze alleen effect zou hebben op kennis van taalregels, expliciete kennis, terwijl voor vloeiend en correct spreken impliciete kennis noodzakelijk is. Het meten van taalvaardigheid blijft een belangrijk onderwerp van discussie. Mede daarom proberen wij in dit proefschrift zoveel mogelijk het leerproces en de leeruitkomsten met elkaar in verband te brengen.

Overzicht van het proefschrift

Hoofdstuk 1: Introductie

Hoofdstuk 1 geeft een algemene introductie van dit proefschrift en laat zien hoe de onderlinge hoofdstukken aan elkaar gerelateerd zijn. Hier worden de drie hoofdzaken beschreven van dit proefschrift met betrekking tot CF-onderzoek en waarvoor het ontwikkelen van GREET noodzakelijk was: het onderzoeken van de rol van CF in het verbeteren van spreekvaardigheid, het onderzoeken van de invloed van individuele verschillen op CF-effect en het onderzoeken van het leerproces naast de leeruitkomsten.
Hoofdstuk 2: Adaptieve correctieve feedback voor het leren van een tweede taal

Hoofdstuk 2 beargumenteert dat een computeromgeving voor taal leren (‘computer-assisted language learning’, CALL) die gebruik maakt van automatische spraakherkenning een optimale omgeving zou kunnen bieden voor CF-onderzoek. CF-onderzoek is gebaat bij experimenteeromgevingen die gecontroleerd zijn, waar individuele verschillen en leergedrag geanalyseerd kunnen worden. In een dergelijke onderzoeksomgeving moeten externe factoren (zoals context, input, gesprekspartner of type CF) consistent zijn zodat herhalingsonderzoek mogelijk is.

In een CALL-systeem is dit mogelijk. De input is constant en de reacties van het systeem (met name de CF) zijn consistent en exact gedefinieerd. Daarnaast kunnen alle interacties van het systeem met de leerder worden opgeslagen. Op deze manier kunnen de leeruitkomsten worden bestudeerd als het resultaat van een taaloefening waarvan precies bekend is wat de leerder heeft gedaan. Hoewel er al verschillende onderzoeken zijn gedaan naar taal leren met de computer, zijn er nog zeer weinig onderzoeken gedaan naar spreekvaardigheid met de computer.

Hoofdstuk 2 laat zien dat gecontroleerd en systematisch onderzoek naar CF nodig is, met meer aandacht voor individuele verschillen in het leren van een tweede taal. Dergelijk onderzoek is mogelijk wanneer er gebruikgemaakt wordt van automatische spraakherkenning in CALL-systemen. Dit hoofdstuk bespreekt de mogelijkheden voor CALL-systemen die individuele spreekkoefeningen aanbieden en de mogelijkheden die dergelijke systemen hebben om adaptief te zijn, door automatische analyse van leergedrag en voorkeuren. In dit hoofdstuk stellen we een experimentopzet voor die haalbaar is vanuit een technologisch, methodologisch en didactisch perspectief.

Hoofdstuk 3: Spreekkoefeningen met feedback op grammatica in een CALL-systeem met automatische spraakherkenning

Hoofdstuk 3 gaat in detail in op de opzet van GREET: hoe de spreekkoefeningen zijn vormgegeven en hoe op basis van automatische spraakherkenning automatisch CF wordt gegeven. Het programma GREET wordt geëvalueerd in een taalleerexperiment met deelnemers met verschillende moedertalen. We vergelijken hier een conditie waarin leerders automatische CF krijgen met een conditie waarin leerders de spreekkoefening doorlopen zonder CF. Analyses van taalvaardigheid, de logdata van de oefening en de evaluaties van de gebruikers laten zien dat GREET succesvol was als spreekkoefening voor leerders van Nederlands als tweede taal. Leerders in beide condities verbeterden hun taalvaardigheid: de spreekkoefening bleek effectief met CF en zonder CF. Wel was er een duidelijk verschil in gedrag van de leerders. De CF-conditie resulteerde in meer gesproken uitingen per leerder en de leerders in de CF-conditie evalueerden het systeem positiever. We concluderen dat GREET als experimentomgeving geschikt is voor
tweedetaalonderzoek naar CF voor gesproken taal leren. In hoofdstuk 3 bespreken we het functioneren van het systeem en de verkregen taalvaardigheidscores en het waargenomen leerdergedrag. Ook verbinden we de resultaten met tweedetaalverwervingstheorieën. We bespreken het vermoeden dat een toegevoegd effect van CF voor taal leren niet zichtbaar was vanwege het hoge taalvaardigheidsniveau van de leerders.

**Hoofdstuk 4: Het effect van correctieve feedback voor het leren van woordvolgorde**

Individuele verschillen tussen leerders kunnen invloed hebben op het effect van CF. Een voorbeeld hiervan is het taalvaardigheidsniveau van de leerder: sommige onderzoeken vonden dat het effect van CF veranderde naarmate de taalvaardigheid van de leerder hoger was. In hoofdstuk 3 leek het taalvaardigheidsniveau van de leerders ook van invloed. Dit vermoeden onderzochten we in hoofdstuk 4 in een gedetailleerd experiment naar de rol van CF voor de ontwikkeling van spreekvaardigheid. Dit experiment is een vervolg op het experiment in hoofdstuk 3, met dezelfde opzet maar met enkele aanpassingen aan GREET voor gebruiksvriendelijkheid, logdata-opslag en aan de pre- en posttests. Naar aanleiding van hoofdstuk 3 werden leerders van een lager taalvaardigheidsniveau gerekruiteerd (op CEF niveau A1-A2). Opnieuw vergeleken we twee condities: een CF- en een NOCF-conditie. Bij de lager taalvaardige leerders vinden we verschillen tussen de condities in leeruitslagen, leergeraad en de evaluatiescore. In de analyse relatieren we het leergeraad aan de leeruitslag. De verschillen in leergeraad en de evaluatie worden besproken in verband met de discussies in de CF-literatuur over de rol van input, output en het corrigeren van fouten.

**Hoofdstuk 5: Spreekoeefeningen in de tweede taal met CALL: Automatische feedback en opleidingsniveau**

Hoofdstuk 5 bespreekt de invloed van het opleidingsniveau. We vergelijken een groep laagopgeleide taalleerders die hebben geoefend met GREET met de hoogopgeleide leerders uit hoofdstuk 4. We rekruteerden laagopgeleide taalleerders op hetzelfde taalvaardigheidsniveau (A1-A2) bij verschillende centra in Nederland zodat opleidingsniveaus verschillen van basisschool tot universitair niveau. We vonden een interactie tussen het niveau van opleiding met de effectiviteit van de oefening en met de CF- en met de NOCF-conditie. De verschillen tussen opleidingsniveaus zijn zichtbaar bij de leeruitslag en het leergeraad. De hoogopgeleide leerders verbeterden hun taalvaardigheid in beide condities, maar dit gold niet voor de laagopgeleide leerders. In het leergeraad waren belangrijke aanwijzingen waarom de oefening voor de ene groep wel, en voor de andere groep niet effectief was. Dit hoofdstuk laat zien dat bevindingen van taalonderzoek met hoogopgeleide leerders niet zomaar te generaliseren zijn naar laagopgeleide leerders.
Hoofdstuk 6: Discussie en conclusies

Hoofdstuk 6 geeft een uitgebreide discussie van de resultaten. We concluderen dat een CALL-systeem met automatische spraakherkenning een belangrijke innovatie is voor tweedetaalverwervingsonderzoek. Een CALL-systeem maakt het mogelijk spreken te oefenen onder gecontroleerde experimentele condities met gedetailleerde logbestanden van het leer(der)gedrag.

Het gedrag van de leerder gedurende oefeningen met een CALL-systeem verschaf belangrijke informatie. Voor onderzoeksdoeleinden geeft het loggen van leerdersgedrag inzicht in individuele variatie in het oefenen en registreert en kwantificeert het de hoeveelheden tweedetaal-input, output en hoe deze veranderen. Deze informatie kan worden gekoppeld aan uitkomsten in taalvaardigheidstoetsen voor meer inzicht in taalleerprocessen.

Het onderzoek in dit proefschrift laat zien dat de effectiviteit van CF genuanceerd moet worden uitgedrukt. Spreekoefeningen met CF blijken effectiever te kunnen zijn dan spreekoefeningen zonder CF, maar dat hangt samen met individuele variabelen. Individuele verschillen tussen leerders hebben invloed op de effectiviteit van CF. De taalvaardigheid van de leerder is een belangrijke factor. Leerders die laag- en middeltaalvaardig zijn hebben meer profijt van oefening met CF dan van gesproken oefening alleen. Daarnaast is de opleidingsachtergrond van de leerder van invloed. We vinden dat bevindingen voor hoogopgeleide leerders afwijken van die voor laagopgeleide leerders. Laagopgeleide leerders hadden geen profijt van de spreekoefening, met of zonder CF.

Toekomstig onderzoek
We hebben laten zien dat een CALL-systeem met automatische spraakherkenning, zoals GREET, effectief is voor onderzoek naar de invloed van CF op spreekvaardigheid en voor het verbeteren van tweedetaal-speektaalvaardigheid in het algemeen. Een CALL-systeem biedt zeer interessante mogelijkheden voor toekomstig onderzoek. Met GREET, of een vergelijkbaar systeem, kan systematisch onderzoek worden gedaan naar de variabelen die invloed hebben op CF: verschillende doelstructuren, CF-typen en individuele verschillen. In longitudinal studies kunnen nieuwe onderzoeksvragen beantwoord worden over de interacties van leergedrag en CF (typen) en hoe deze veranderen. Op basis van analyses van leergedrag zou het systeem zich bovendien automatisch kunnen aanpassen aan de leerder. Met betrekking tot het analyseren van leergedrag zou GREET nog kunnen worden uitgerust met mogelijkheden om oogbewegingen te volgen (‘eye tracking’), wat inmiddels compatibel is met een laptop. Dit zou weer een extra laag aan informatie geven over hoe de leerder het scherm leest alvorens de vragen te beantwoorden. Zo kan worden nagegaan of er tekenen zijn van het opmerken van de specifieke vorm (‘noticing’) bij het krijgen van CF. Het kan ook interessant zijn om de relatie van CF-voorkeur en CF-effectiviteit te onderzoeken. Dat kan met behulp van gedetailleerde logbestanden en tussentijdse vragen over motivatie. Een andere
interesting possibility is the development of serious games with exercises where effective CF is given, while the learner remains motivated to practice a lot.

Our experiments with GREET were mainly done outside the learners' courses. It is easy to imagine that exercises such as those of GREET would be part of a language course. Certainly, for speaking practice, where much repetition (and CF needed) is required, CALL systems with automatic speech recognition can be very valuable for (second)language learning. The exercise data of the learners could be used to conduct (second)language acquisition research and to improve the design of the CALL system. In this way, a CALL system can not only generate new findings for (second)language acquisition research but also data to continuously improve itself.
Bart Penning de Vries studied at Utrecht University, where he obtained a Bachelor of Arts degree in English Language and Culture in 2007, and a Master of Arts degree in Linguistics in 2009. His MA thesis was a study on the role of Dutch phonotactics in speech segmentation, based on his experiments with an artificial language. After working as a research assistant at the Utrecht institute for Linguistics (UiL-OTS), he started his doctoral research in Nijmegen on January 1, 2010.

This dissertation results from doctoral research conducted at the Centre for language and Speech technology at the Radboud University Nijmegen.

He is currently working at the research department of Bureau-ICE on the development of (language) tests for Dutch education.