Bart Penning de Vries*, Catia Cucchiariini, Stephen Bodnar,
Helmer Strik and Roeland van Hout

Effect of corrective feedback for learning verb second

DOI 10.1515/iral-2016-0121

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.

***Corresponding author: Bart Penning de Vries**, Centre for Language Studies, Radboud University Nijmegen, E-mail: bardtpdv@gmail.com

Catia Cucchiariini: E-mail: c.cucchiariini@let.ru.nl, **Stephen Bodnar:** E-mail: sbodnar@gmail.com, **Helmer Strik:** E-mail: w.strik@let.ru.nl, **Roeland van Hout:** E-mail: r.vanhout@let.ru.nl, Centre for Language Studies, Radboud University Nijmegen

Keywords: Computer-assisted language learning (CALL), corrective feedback (CF), second language acquisition (SLA), automatic speech recognition (ASR)

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 and Spada 2006; Lyster and 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 and 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 and Ranta 1997; Lyster 2004), and (b) learner-internal variables: individual differences such as learner proficiency (Mackey and 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 and Spada 2006; Ellis 2010; Ellis and Sheen 2006; Lyster et al. 2013; Goo and Mackey 2013). The focus in CF research, therefore, is shifting from whether CF is effective, to what determines its effectiveness (Mackey and 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 et al. 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 et al. 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 and 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 and 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 and Long 2003). This adaptation is desirable for pedagogy, and CALL systems make it feasible (Chapelle 2007; Heift and 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 (Section 2.1), and studies that investigate how learner characteristics, specifically learner proficiency, mediate CF effectiveness (Section 2.2), and how learner behaviour in CALL exercises can be recorded and related to language learning outcomes and processes (Section 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 and 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 and Spada 2006; Hegelheimer and 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 and Spada 2006; e. g. see the discussion between Lyster and Ranta [2013] and Goo and 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 et al. 2003). Moreover, individual learners vary in the ways they react to, and benefit from interactional CF (e. g. Mackey 1999; Mackey et al. 2002, 2003; Sheen 2004, 2008; Swain and Lapkin 1998). This suggests that CF may be individualized to improve learning results (cf. 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 and 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 and Lantolf 1994; Mackey and 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 and 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 and 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. Diéz-Bedmar and 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 and Ranta 1997; Nicholas et al. 2001).

Indirect inferences about noticing can also be made through investigation of logged data in a CALL context (Hegelheimer and 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 et al. 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.

Table 1: The languages spoken by the participants, and their corresponding word order in the main clause.

Word order	Language knowledge of the Participants	
	L1	L2(s) besides Dutch
VSO	Arabic, Tagalog	Arabic
SOV	Chinese (2), French (2), Hungarian (2), Japanese, Pashto, Persian (3)	Japanese, Latin, Persian, Urdu
SVO	Czech, German (2)*, Italian, Lithuanian, Polish (2), Russian (3), Spanish (5), <i>Temne</i> , Thai (2)	Danish*, English (28), French (8), German (6)*, Indonesian, <i>Krio</i> , Russian (2), Spanish (3)

Source: <http://aboutworldlanguages.com/> except those in italics.
Note: The asterisk * indicates that the language has verb second in the main clause.

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 min to complete. After the proficiency tests, the participant moved on to 45 min 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 min) 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 min.

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 and 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 67 questions (25 target, 42 distracters); the second session contains 60 questions (19 targets, 41 distracters). Learners vary in their speed of practice, and some complete all questions of a session within the 45 min of practice. To control for time-on-task, learners were automatically restarted at the beginning to redo questions to complete 45 min 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.

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

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.

Target structure			
(1) Subject-initial main clause			
Jan	koopt	morgen	bloemen
<i>Jan</i>	<i>buys (3SG)</i>	<i>tomorrow</i>	<i>flowers</i>
Subject	Verb	Adverbial	Object
(2) Adverbial-intial sentence: Postverbal subject			
Morgen	koopt	Jan	Bloemen
<i>Tomorrow</i>	<i>buys (3SG)</i>	<i>Jan</i>	<i>Flowers</i>
Adverbial	Verb	Subject	Object

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 and Selinker 2008).

3.4.2 Corrective feedback in our CALL system

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

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

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 and 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 s. If the participant did not record an answer within that time, the response was scored as incorrect.

Example 1. DCT target item

Question:

Wat doet Kim om twaalf uur?

[Illustration: photo of sandwich]

Om twaalf uur...

boterham + eten

(‘What does Kim do at twelve o’ clock?/ At twelve ... / 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 s (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 A 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 (Sections 4.1 and 4.2), followed by an exploration of other differences between the conditions (language background [Section 4.3], post-test questionnaire [Section 4.4]). Then we proceed with an analysis of logged data of learner behaviour (Section 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=0.960; scores above 0.7 are generally considered reliable). Figure 2 shows the individual trajectories of the DCT scores

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.

split out by treatment condition. The DCT pre-test means are (M (CF)=0.39, SD =0.34, range: 0.00–0.94; M (NOCF)=0.30, SD =0.32, range: 0.00–0.94). There is no difference between the groups at pre-test (F (1,30)=0.577, p =0.453), 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)=0.29, SD =0.30; M (A2)=0.46, SD =0.37). However, within both CEFR levels the individual V2 proficiency scores still range from zero knowledge to full competence (range: 0.00–0.94).

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 Sections 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)=0.579, p =0.455).

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.

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).

	DCT mean scores			
	Complete group		Trimmed group	
	pre M (SD)	post M (SD)	Pre M (SD)	post M (SD)
CF	0.39 (0.344)	0.53 (0.295)	0.22 (0.198)	0.42 (0.254)
NOCF	0.30 (0.318)	0.37 (0.347)	0.17 (0.169)	0.24 (0.240)

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

Effect	ANOVA of DCT scores	
	Complete group (N = 31)	Trimmed group (N = 24)
Time	$F(1,29) = 13.912, p = 0.001,$ $\eta^2_p = 0.324$	$F(1,22) = 2.429, p = 0.000,$ $\eta^2_p = 0.481$
Condition	$F(1,29) = 1.216, p = 0.279,$ $\eta^2_p = 0.040$	$F(1,22) = 2.042, p = 0.167,$ $\eta^2_p = 0.085$
Time by condition	$F(1,29) = 1.515, p = 0.228,$ $\eta^2_p = 0.050$	$F(1,22) = 4.352, p = 0.049,$ $\eta^2_p = 0.165$

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 ($F(1,15) = 10.041, p = 0.006, \eta^2 = 0.401$; trimmed $F(1,11) = 18.470, p = 0.001, \eta^2 = 0.627$), and for the NOCF condition the effect of time is not significant ($F(1,15) = 4.232, p = 0.059, \eta^2 = 0.232$; trimmed: $F(1,11) = 3.617, p = 0.084, \eta^2 = 0.247$). 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 = 0.752). For the fillers, the reliability is lower (GJT fillers is 40, Cronbach’s Alpha = 0.559). 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 $F(1, 30) = 1.237, p = 0.275$; trimmed: $F(1, 23) = 1.817, p = 0.191$. For the filler items: $F(1, 30) = 3.067, p = 0.090$; trimmed: $F(1, 23) = 2.984, p = 0.098$).

The GJT scores are correlated with the DCT scores (pre-test scores: $r = 0.585$, $p = 0.001$; trimmed group: $r = 0.225$, $p = 0.291$; post-test scores: $r = 0.621$, $p = 0.000$; trimmed group: $r = 0.553$, $p = 0.005$). In interpreting this correlation, we need to take into account that the GJT is a receptive, written, forced choice test.¹ 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 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 $N = 12$).

		GJT mean scores			
		Complete group		Trimmed group	
		Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)	Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)
Targets	CF	0.56 (0.15)	0.65 (0.17)	0.51 (0.12)	0.60 (0.16)
	NOCF	0.49 (0.17)	0.55 (0.20)	0.45 (0.12)	0.53 (0.20)
Fillers	CF	0.65 (0.09)	0.64 (0.13)	0.63 (0.09)	0.61 (0.13)
	NOCF	0.59 (0.10)	0.63 (0.13)	0.58 (0.07)	0.59 (0.08)

Table 6: Results of the GJT in a 2×2 mixed factorial ANOVA, for both filler and target items.

ANOVA of GJT scores			
Effect		Complete ($n = 31$)	Trimmed ($n = 24$)
Targets	Time	$F(1,29) = 4.812$, $p = 0.036$, $\eta^2_p = 0.142$	$F(1,22) = 3.656$, $p = 0.069$, $\eta^2_p = 0.143$
	Condition	$F(1,29) = 2.679$, $p = 0.112$, $\eta^2_p = 0.085$	$F(1,22) = 2.435$, $p = 0.133$, $\eta^2_p = 0.100$
	Time by condition	$F(1,29) = 0.266$, $p = 0.610$, $\eta^2_p = 0.009$	$F(1,22) = 0.009$, $p = 0.925$, $\eta^2_p = 0.000$
Fillers	Time	$F(1,29) = 0.471$, $p = 0.498$, $\eta^2_p = 0.016$	$F(1,22) = 0.006$, $p = 0.937$, $\eta^2_p = 0.000$
	Condition	$F(1,29) = 1.023$, $p = 0.320$, $\eta^2_p = 0.034$	$F(1,22) = 1.497$, $p = 0.234$, $\eta^2_p = 0.064$
	Time by condition	$F(1,22) = 0.009$, $p = 0.925$, $\eta^2_p = 0.000$	$F(1,22) = 0.524$, $p = 0.477$, $\eta^2_p = 0.023$

1 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 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 = 0.744) 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 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).

Post-test questionnaire: Evaluation scores		
	Complete group	Trimmed group
	<i>M (SD); range</i>	<i>M (SD); range</i>
CF	4.10 (0.43); 3.3–4.7	4.03 (0.44); 3.3–4.7
NOCF	3.76 (0.64); 2.8–5.0	3.51 (0.42); 2.8–4.1

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 ($F(1,29) = 3.058$, $p = 0.091$), but the difference is significant for the trimmed group ($F(1,22) = 8.577$, $p = 0.008$). 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 ($r = 0.760$, $p = 0.001$), but not for the CF group ($r = 0.262$, $p = 0.326$); 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 ($F(1,22) = 6.401$, $p = 0.019$) and Question 6 in the complete and trimmed group ($F(1,29) = 5.194$, $p = 0.030$ and $F(1,22) = 5.852$, $p = 0.024$). 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,

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

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.

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 (Section 4.5.1), the number of attempts to answer the questions and how that changed over time (Section 4.5.2), and the number of times learners produced a correct answer and how the accuracy of the attempts changed over time (Section 4.5.3).

4.5.1 Questions practiced

The treatment session contains 127 questions, of which 44 are target questions. If learners complete all questions, they restart and continue until 45 min of

practice time is up (referred to as ‘looping’). Table 8 gives the number of questions practiced during 90 min 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 8: Means and SDs of the number of questions (all and target only) in 90 min of GREET treatment, CF vs. NOCF, for the complete and the trimmed group.

	Questions practiced			
	Complete group		Trimmed group	
	All <i>M (SD)</i>	Targets <i>M (SD); range</i>	All <i>M (SD)</i>	Targets <i>M (SD); range</i>
CF	142.94 (42.95)	47.56 (15.40); 27–81	126.00 (31.48)	41.33 (10.44); 27–60
NOCF	141.4 (27.55)	47.13 (9.58); 32–66	131.08 (25.40)	43.50 (6.40); 32–51

Table 8 shows that the number of target questions practiced is similar for both groups (complete: $F(1,29) = 0.009$, $p = 0.927$; trimmed: $F(1,22) = 0.376$, $p = 0.546$). Figure 4 plots the mean number of questions practiced against DCT pre-test proficiency score.

The pre-test scores are correlated to the number of questions practiced (CF (complete): $r = 0.830$, $p = 0.000$; CF(trimmed): $r = 0.687$, $p = 0.013$; NOCF(complete): $r = 0.706$, $p = 0.003$, NOCF(trimmed): $r = 0.126$, $p = 0.696$). 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

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

	Mean number of attempts, and attempts per question			
	Complete group		Trimmed group	
	Attempts	ApQ	Attempts	ApQ
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
CF	84.81 (21.65)	1.93 (0.724)	84.50 (22.34)	2.14 (0.715)
NOCF	55.40 (18.84)	1.17 (0.311)	50.83 (17.82)	1.16 (0.336)

Figure 5: Scatterplot of the mean number of attempts per question in 90 min 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.² As a result, the number of ApQ for NOCF is not correlated with V2 proficiency level

² One NOCF participant practices more (2.13 ApQ) than all other NOCF participants (approx 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.

($r = -0.063$, $p = 0.824$). For the CF group, there is an inverse correlation of ApQ with V2 proficiency ($r = -0.653$, $p = 0.006$): 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.³

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 ($\pm 2SE$) 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($SD = 0.87$) to 1.97 ($SD = 0.58$). 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 hint (under 2 ApQ). In an ANOVA we find that the decrease of ApQ is significant (Effect of time: $F(1,29) = 5.690$, $p = 0.024$, $\eta^2_p = 0.164$; trimmed $F(1,22) = 6.174$, $p = 0.021$, $\eta^2_p = 0.219$), and we also find a difference between the groups over time (Interaction of time by condition: $F(1,29) = 7.490$, $p = 0.010$, $\eta^2_p = 0.205$; trimmed $F(1,22) = 8.247$, $p = 0.009$, $\eta^2_p = 0.273$). 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.

³ 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 min interval his ApQ changes from 4.5, 5, 4, 3, 3.7, to 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).

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

	Number of OKs	
	Complete	Trimmed
	<i>M (SD)</i>	<i>M (SD)</i>
CF	44.75 (16.72)	38.08 (12.09)
NOCF	19.07 (2.43)	10.33 (8.86)

The number of OKs for the NOCF condition is significantly lower (complete: $F(1,29) = 14.757$, $p = 0.001$; trimmed: $F(1,22) = 41.127$, $p = 0.000$). 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

⁴ The CF participants cannot retry after they see an OK. Several participants have indicated that they like that the system allows them to practice at their own speed – however, in our latest experiment, two participants remarked that the speed of the feedback was ‘too high’. This seems to suggest that automatically moving them forward is not the best option, per se. For experiment purposes, though, we have had to make this non-optional.

restarting an utterance).⁵ The trimmed group has a lower number of OKs which is self-evident as the ceiling learners were removed.

Figure 7 shows that for the CF and NOCF condition, the number of OKs correlates with the initial V2 proficiency level (CF: $r = 0.889$, $p = 0.000$, NOCF: $r = 0.884$, $p = 0.000$). In addition, we see that the difference between the CF and NOCF group in number of OKs becomes smaller as pre-test proficiency increases.

Figure 7: Scatterplot showing the number of correctly formed sentences (OKs) during 90 min of treatment per participant, against pre-test proficiency.

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 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 $F(1,29) = 4.294$, $p = 0.047$,

⁵ 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 8: Bar graphs with standard errors ($\pm 2SE$) showing the ratio of OKs per attempt per 45-minute session for the complete and the trimmed group.

$\eta^2_p = 0.129$; trimmed $F(1,22) = 5.985$, $p = 0.023$, $\eta^2_p = 0.214$). There is no difference between the groups in the change over time (Interaction time by condition: complete $F(1,29) = 1.533$, $p = 0.226$, $\eta^2_p = 0.050$; trimmed $F(1,22) = 0.543$, $p = 0.469$, $\eta^2_p = 0.024$). 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 min 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? (RQ.1)

Learners practiced for 90 min 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 and 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; Kartchava and 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 evidence of transfer of the V2 target structure from the L1 (cf. Ringbom 1992), and we also found evidence to suggest transfer from L2 German. We did not find other influences of the 17 different L1s that were present in our participant sample.

5.2 To what extent does learner behaviour differ between the treatment conditions?(RQ.2)

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.

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 and 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? (RQ.3)

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 and Tower 2004; Heift and 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 and 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 and Ranta 2013; Goo and 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 (Doughty and Long 2003), and adaptation of CALL systems to learners as an attractive and feasible future direction for CALL (Chapelle 2007; Heift and 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.

Acknowledgements: We would like to thank two anonymous reviewers for their useful comments.

References

- Aljaafreh, A. & J. Lantolf. 1994. Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal* 78(4). 465–483. doi:10.1111/j.1540-4781.1994.tb02064.x/abstract.
- Ammar, A. & N. Spada. 2006. One size fits all?: Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition* 28(4). 543–574. doi:10.1017/S0272263106060268.
- Bohnacker, U. 2006. When Swedes begin to learn German: From V2 to V2. *Second Language Research* 22(4). 443–486. doi:10.1191/0267658306sr2750a.
- Chapelle, C. 2005. Interactionist SLA theory in CALL research. In J. Egbert & G. Petrie (eds.), *Research perspectives on CALL*, 53–64. Mahwah, NJ: Erlbaum.
- Chapelle, C. 2007. Technology and second language acquisition. *Annual Review of Applied Linguistics* 27. 98–114. doi:10.1017/S0267190508070050.
- Collentine, J. 2000. Insights into the construction of grammatical knowledge provided by user-behavior tracking technologies. *Language Learning & Technology* 3(2). 44–57. Retrieved from <http://llt.msu.edu/vol3num2/collentine/>.
- De Bot, K. 1996. Review article: The psycholinguistics of the output hypothesis. *Language Learning* 46(3). 529–555.
- De Bot, K., W. Lowie & M. Verspoor. 2007. A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1). 7–21. doi:10.1017/S1366728906002732.
- De Graaff, R. 1997. The esperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition* 19(2). 249–276.
- DeKeyser, R. 1998. Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (eds.), *Focus on form in classroom language acquisition*, 42–63. New York: Cambridge University Press.
- DeKeyser, R. 2007. *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, UK: Cambridge University Press.
- Diez-Bedmar, M. & P. Perez-Paredes. 2012. The types and effects of peer native speakers' feedback on CMC. *Language Learning & Technology* 16(1). 62–90. Retrieved from <http://eric.ed.gov/?id=EJ972345>.
- Doughty, C. & M. Long. 2003. Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology* 7(3). 50–75. Retrieved from <http://ir.nul.nagoya-u.ac.jp/jspui/handle/2237/6283>.
- Ehrman, M., B. Leaver & R. Oxford. 2003. A brief overview of individual differences in second language learning. *System* 31(3). 313–330. doi:10.1016/S0346-251X(03)00045-9.
- Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

- Ellis, R. 2000. *Learning a second language through interaction. Studies in bilingualism*, 17. Philadelphia: John Benjamins.
- Ellis, N. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 141–188.
- Ellis, R. 2005. Measuring implicit and explicit knowledge of a second language: A Psychometric Study. *Studies in Second Language Acquisition* 27(2). 141–172.
- Ellis, R. 2007. The differential effects of corrective feedback on two grammatical structures. In A. Mackey (ed.), *Conversational interaction in second language acquisition: A series of empirical studies*, 339–360. Oxford: Oxford University Press.
- Ellis, R. 2009. The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics* 30. 474–509.
- Ellis, R. 2010. Epilogue. *Studies in Second Language Acquisition* 32(2). 335–349. doi:10.1017/S0272263109990544.
- Ellis, R., H. Basturkmen & S. Loewen. 2001. Learner uptake in communicative ESL lessons. *Language Learning* 51(2). 281–318. doi:10.1111/1467-9922.00156.
- Ellis, R., S. Loewen & R. Erlam. 2006. Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition* 28(2). 339–368.
- Ellis, R. & Y. Sheen. 2006. Reexamining the role of recasts in second language acquisition. *Studies in Second Language Acquisition* 28(04). 575–600. doi:10.1017/S027226310606027X.
- Fischer, R. 2007. How do we know what students are actually doing? Monitoring students' behavior in CALL. *Computer Assisted Language Learning* 20(5). 409–442. doi:10.1080/09588220701746013.
- Gass, S. 1997. *Input, Interaction, and the Second Language Learner*. Mahwah, NJ: Lawrence Erlbaum.
- Gass, S. 2003. Input and interaction. In C. Doughty & M. Long (eds.), *The handbook of second language acquisition*. 224–255. MA: Blackwell Publishing.
- Gass, S. & L. Selinker. 2008. Second language acquisition: An introductory course. *Language* 71. 1–7. doi: 10.2307/416225.
- Goo, J. & A. Mackey. 2013. The case against the case against recasts. *Studies in Second Language Acquisition* 35(01). 127–165. doi:10.1017/S0272263112000708.
- Han, Z. 2001. Fine-tuning corrective feedback. *Foreign Language Annals* 34(6). 582–599.
- Hanaoka, O. 2007. Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research* 11. 459–479. doi:10.1177/1362168807080963.
- Hattie, J. & H. Timperley. 2007. The power of feedback. *Review of Educational Research* 77(1). 81–112. doi: 10.3102/003465430298487.
- Hegelheimer, V. & C. Chapelle. 2000. Methodological issues in research on learner-computer interactions in CALL. *Language Learning & Technology* 4(1). 41–59. Retrieved from <http://llt.msu.edu/vol4num1/hegchap/default.html>.
- Hegelheimer, V. & D. Tower. 2004. Using CALL in the classroom: Analyzing student interactions in an authentic classroom. *System* 32(2). 185–205. doi:10.1016/j.system.2003.11.007.
- Heift, T. 2002. Learner control and error correction in ICALL: Browsers, peekers, and adamantans. *CALICO Journal* 19(2). 295–313.
- Heift, T. 2004. Corrective feedback and learner uptake in CALL. *ReCALL* 16(02). 416–431. doi:10.1017/S0958344004001120.

- Heift, T. 2010. Prompting in CALL: A longitudinal study of learner uptake. *Modern Language Journal* 94(2). 198–216 (8).
- Heift, T. & A. Rimrott. 2012. Task-related variation in computer-assisted language learning. *The Modern Language Journal* 96(4). 525–543. doi:10.1111/j.1540-4781.2012.01392.x.
- Heift, T. & M. Schulze. 2007. *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York: Routledge.
- Hulstijn, J. 1997. Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition* 19. 131–143.
- Hulstijn, J. 2000. The use of computer technology in experimental studies of second language acquisition: A survey of some techniques and some ongoing studies. *Language Learning & Technology* 3(2). 32–43. Retrieved from <http://dare.uva.nl/record/84048>.
- Iwashita, N. 2003. Negative feedback and positive evidence in task-based interaction: Differential effects on L2 development. *Studies in Second Language Acquisition* 25(1). 1–36.
- Jean, G. & D. Simard. 2011. Grammar teaching and learning in L2: Necessary, but boring?. *Foreign Language Annals* 44(3). 467–494. doi:10.1111/j.1944-9720.2011.01143.x.
- Jordens, P. 1988. The acquisition of word order in Dutch and German as L1 and L2. *Second Language Research* 4(1). 41–65. doi:10.1177/026765838800400103.
- Kartchava, E. & A. Ammar. 2013. Noticing and learning: Relationship patterns. *Studies in English Language Teaching* 1(1). 8–25. Retrieved from <http://www.scholink.org/ojs/index.php/selt/article/download/63/69>.
- Li, S. 2010. The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning* 60(2). 309–365. doi:10.1111/j.1467-9922.2010.00561.x.
- Li, S. 2013. The interface between feedback type, L2 proficiency, and the nature of the linguistic target. *Language Teaching Research*. 1–24. doi:10.1177/1362168813510384.
- Lin, Y. & J. Hedgcock. 1996. Negative feedback incorporation among high-proficiency and low-proficiency Chinese-speaking learners of Spanish. *Language Learning* 46(4). 567– 611.
- Loewen, S. 2004. Uptake in incidental focus on form in meaning-focused ESL lessons. *Language Learning* 54(1). 153–188. doi:10.1111/j.1467-9922.2004.00251.x/abstract.
- Long, M. 1996. The role of linguistic environment in second language instruction. In Ritchie, W. & T. Bhatia (eds.), *Handbook of second language Acquisition* 2. 413–468. New York: Academic Press.
- Lyster, R. 2004. Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition* 26(3). 399–432. doi:10.1017/S0272263104263021.
- Lyster, R. & L. Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition* 20. 37–66.
- Lyster, R. & L. Ranta. 2013. Counterpoint piece: The case for variety in corrective feedback research. *Studies in Second Language Acquisition* 35(01). 167–184. doi:10.1017/S027226311200071X.
- Lyster, R. & K. Saito. 2010. Interactional feedback as instructional input: A synthesis of classroom SLA research. *Language, Interaction and Acquisition* 1(2). 276–297. doi:10.1075/lia.1.2.07lys.
- Lyster, R., K. Saito & M. Sato. 2013. Oral corrective feedback in second language classrooms. *Language Teaching* 46(1). 1–40. doi:10.1017/S0261444812000365.
- Mackey, A. 1999. Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition* 21. 557–587.
- Mackey, A. 2006. Epilogue: From introspections, brain scans, and memory tests to the role of social context: Advancing research on interaction and learning. *Studies in Second Language Acquisition* 28(2). 369–379. doi:10.1017/S0272263106060153.

- Mackey, A. & S. Gass. 2006. Introduction. *Studies in Second Language Acquisition* 28(2). 169–178. doi:10.1017/S0272263106060086.
- Mackey, A., R. Oliver & J. Leeman. 2003. Interactional input and the incorporation of feedback: An exploration of NS-NNS and NNS-NNS adult and child dyads. *Language Learning* 53(1). 35–66.
- Mackey, A. & J. Philp. 1998. Conversational interaction and second language development: Recasts, responses, and red herrings?. *The Modern Language Journal* 82. 338–356. doi: 10.1111/j.1540-4781.1998.tb01211.x.
- Mackey, A., J. Philp, A. Fujii, T. Egi & T. Tatsumi. 2002. Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (ed.), *Individual differences and instructed language learning*, 181–208. Amsterdam: John Benjamins.
- Malmberg. 1999. *Nieuwe Buren: Opzoekboek*. Den Bosch: Malberg.
- Nicholas, H., P. Lightbown & N. Spada. 2001. Recasts as feedback to language learners. *Language Learning* 51(4). 719–758. doi:10.1111/0023-8333.00172/abstract.
- Norris, J. & L. Ortega. 2003. Defining and measuring SLA. In C. Doughty & M. Long (eds.), *The handbook of second language acquisition*, 717–761. Malden: Blackwell.
- Odlin, T. 1989. *Language transfer*. Cambridge: Cambridge University Press.
- Penning de Vries, B., C. Cucchiari, S. Bodnar, H. Strik & R. van Hout. 2014. Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning*. 1–27. doi:10.1080/09588221.2014.889713.
- Penning de Vries, B., C. Cucchiari, H. Strik & R. van Hout. 2011. Adaptive corrective feedback in second language learning. In S. de Wannemacker, G. Clarebout, & P. de Causmaecker (eds.), *Interdisciplinary approaches to adaptive learning: A look at the neighbours*, Vol. 126, 1–14. Heidelberg: Springer.
- Philp, J. 2003. Constraints on “noticing the gap”: Nonnative speakers’ noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition* 25(1). 99–126.
- Put, J. & J. Peekel. 2001. *Nieuwe Buren*. Den Bosch: Malmberg.
- Ringbom, H. 1992. On L1 transfer in L2 comprehension and L2 production. *Language Learning* 42(1). 85–112.
- Robinson, P. 2003. The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies* 21(2). 45–105.
- Russell, J. & N. Spada. 2006. The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. Norris & L. Ortega (eds.), *Synthesizing Research on Language Learning and Teaching*, 133–164. Amsterdam: Benjamins.
- Sanz, C. & K. Morgan-Short. 2004. Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning* 54(1). 35–78. doi:10.1111/j.1467-9922.2004.00248.x/abstract.
- Sauro, S. 2009. Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology* 13(1). 96–120.
- Schmidt, R. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11. 129–158. doi:10.1093/applin/11.2.129.
- Schmidt, R. 1993. Awareness and second language acquisition. *Annual Review of Applied Linguistics* 13. 206–226.
- Schmidt, R. 1995. Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. W. Schmidt (ed.), *Attention and awareness in foreign language learning*, 1–63. Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.

- Schulz, R. 1996. Focus on form in the foreign language classroom: Students' and teachers' views on error correction and the role of grammar. *Foreign Language Annals* 29(3). 343–364. doi:10.1111/j.1944-9720.1996.tb01247.x.
- Schulze, M. 2011. CALL in the year 2011. Still in search of research paradigms? Presentation at EuroCALL 2011, Nottingham.
- Sheen, Y. 2004. Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research* 8(3). 263–300.
- Sheen, Y. 2008. Recasts, language anxiety, modified output, and L2 learning. *Language Learning* 58(4). 835–874. doi:10.1111/j.1467-9922.2008.00480.x/full.
- Sheen, Y. 2010. Introduction: The role of oral and written corrective feedback in SLA. *Studies in Second Language Acquisition* 32(2). 169–179. doi:10.1017/S0272263109990489.
- Shintani, N., R. Ellis & W. Suzuki. 2014. Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning* 64(1). 103–131. doi:10.1111/lang.12029.
- Skehan, P. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Smith, B. 2012. Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology* 16(3). 53–81. Retrieved from <http://llt.msu.edu/issues/october2012/smith.pdf>.
- Swain, M. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, & C. Madden (eds.), *Input and second language acquisition*, 235–256. Rowley, MA: Newbury House.
- Swain, M. 1995. Three functions of output in second language learning. In G. Cook & B. Seidlhofer (eds.), *Principle and practice in applied linguistics: Studies in honor of William E. Rutherford*, 125–144. Oxford: Oxford University press.
- Swain, M. & S. Lapkin. 1998. Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal* 82(3). 320–337. doi:10.1111/j.1540-4781.1998.tb01209.x.
- Uzum, B. 2010. Who gains more? A case of motivation and corrective feedback in ESL classes. In *Michigan Teachers of English to Speakers of Other Languages Conferences*, 19–39.
- Van de Craats, I. 2009. The role of 'is' in the acquisition of finiteness by adult learners of Dutch. *Studies in Second Language Acquisition* 31. 59–92. doi:10.1017/S0272263109090032.
- Van der Linden, E. 1993. Does feedback enhance computer-assisted language learning?. *Computers & Education* 21(1/2). 61–65. doi:10.1016/0360-1315(93)90048-N.
- Van Doremalen, J., C. Cucchiariini & H. Strik. 2010. Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*. 1–13. doi:10.1155/2010/973954.
- VanPatten, B. 2004. On the role(s) of input and output in making form-meaning connections. In B. VanPatten, J. Williams, S. Rott & M. Overstreet (eds.), *Form-meaning connections in second language acquisition*, 29–47. Mahwah, NJ: Erlbaum.
- Warschauer, M. 1996. Comparing face-to-face and electronic communication in the second language classroom. *CALICO Journal* 13(2). 7–26.
- Yilmaz, Y. 2012. The relative effects of explicit correction and recasts on two target structures via two communication modes. *Language Learning* 62(4). 1134–1169. doi:10.1111/j.1467-9922.2012.00726.x.
- Young-Scholten, M. 2013. Low-educated immigrants and the social relevance of second language acquisition research. *Second Language Research* 29(4). 441–454. doi:10.1177/0267658313491266.