Tools that Detectives Use: In Search of Learner-related Determinants for Usage of Optional Feedback in a Written Murder Mystery

FREDERIK CORNILLIE

RUBEN LAGATIE

MIEKE VANDEWAEETERE

GERALDINE CLAREBOUT

PIET DESMET

ITEC – iMinds – KU Leuven

ABSTRACT

This paper seeks to identify individual difference factors as determinants of usage of optional metalinguistic corrective feedback (CF) in a written and task-based tutorial CALL environment for English grammar practice that contained gaming features. Previous research in CALL has highlighted the importance of prior knowledge for explaining learners’ usage of CF options (Brandl, 1995; Heift, 2002), but the contribution of meta-cognitive and motivational variables to usage of CF remains unexplored. Based on insights from the literature on tool use (e.g. Clarebout & Elen, 2009), this pilot study (N=36) considered that learners’ usage of optional CF in CALL might, in addition to prior knowledge, be determined by the perceived usefulness of CF and by learners’ achievement goal orientation. Quantitative analysis of tracking and logging data in combination with questionnaire and language test data showed that usage of optional metalinguistic CF was associated with prior explicit L2 knowledge, but no relation was found with perceived usefulness and achievement goal orientation. Future research could benefit from fine-tuning the questionnaires used in this study, as well as from more qualitative in-depth analyses of learners’ perceptions and motives. Also, in future studies game-like features could be implemented in different experimental conditions in order to investigate effects on learner behavior.
INTRODUCTION

Across a wide range of theories in the field of second language acquisition (SLA) and in second language (L2) pedagogy, feedback is increasingly being considered a developmentally useful feature of instructed L2 environments (for reviews see Lyster & Saito, 2010; Russell & Spada, 2006). Specifically, the notion of *corrective feedback* (CF), which may be succinctly defined as any utterance that is intended to correct a learner’s erroneous response (for more comprehensive definitions and theoretical discussion see Carroll, 2001; Ellis, Loewen, & Erlam, 2006), has received substantive and intensified attention in SLA research, especially for the learning of grammar-related features. Although SLA theories are rather divided on the question of whether and how CF facilitates learning, research has altogether been guided by Steven Pinker’s (1989, pp. 9-14) argument that CF can in principle support language development if the following conditions are met: 1) be available in the learner’s environment, 2) be useful (i.e., psycholinguistically relevant), 3) be actually used by language learners, and 4) be “necessary” (i.e., the only feature that explains a specific change in L2 development).

Pinker’s first condition can be satisfied particularly in tutorial computer-assisted language learning (CALL) environments (Hubbard & Bradin Siskin, 2004), which can – notwithstanding the technological and pedagogical challenges involved – provide the learner with immediate, consistent, and error-specific feedback, possibly accompanied by additional help such as extended explanations on the nature of errors. However, the mere availability of such features in CALL programs does not imply that learners will actually use it (Fischer, 2007). Markedly, this is the case for support devices that are non-embedded and hence optional – i.e., the learner needs to click a button to get access to these devices – which are also known as *tools* in the more general research on computer-assisted learning (Clarebout & Elen, 2006). This research recognizes that learners do not always make the right choices for their learning, and argues for the investigation of the complex interplay of factors that might determine tool usage, viz. the nature of the tool, task characteristics, and learner-related factors, such as the motivation to work with particular tools and the functionality that learners attribute to these tools (Clarebout & Elen, 2009). Outside educational settings, usage of technologies can be explained to a significant extent by how these technologies are perceived in terms of usefulness (Davis, 1989), and this line of reasoning seems well applicable to educational research which presupposes that learners’ perceptions of instructional features mediate learner behavior and learning processes (Winne, 2004).

This paper reports on a pilot study in a task-based tutorial CALL environment that includes gaming features and in which CF was available for responses that deviated from the predicted correct responses. The practical aim of the pilot study was to prepare the learning environment and other instruments for a longitudinal experiment. More importantly, the study was intended to explore whether learners actually made use of optional CF, and whether this usage was related to the perceived usefulness of the CF, to learners’ explicit L2 knowledge, or, taking into
account the achievement-oriented nature of gaming ecologies, to their achievement goal orientation. The instruments used in this exploratory study include, first, log files of the learners’ interactions in the software, such as their usage of optional CF, second, tests to assess explicit L2 knowledge and third, questionnaires to measure perceived usefulness of CF and achievement goal orientation. Before turning to the pilot study, we will review two areas of research that are pertinent to the study: research on the perceived usefulness and usage of CF in instructed L2 environments (including CALL research), and the literature on tool use in computer-assisted learning environments.

BACKGROUND RESEARCH

Perceived Usefulness and Usage of CF in Instructed L2 Settings

In instructed L2 settings, there is ample evidence that language learners find CF generally helpful in a wide range of tasks. This finding applies both to feedback given directly by teachers and native speakers (e.g., Chenoweth, Day, Chun, & Luppescu, 1983; Radecki & Swales, 1988; Schulz, 2001) and to feedback generated by or mediated through technology (Cornillie, Clarebout, & Desmet, 2012; Nagata, 1993). Learners have also been found to prefer feedback that comprises metalinguistic explanations rather than less informative ‘correct/incorrect’ feedback (Nagata, 1993) or recasts (correct reformulations of erroneous utterances) (Kim & Mathes, 2001). In addition, research on meaning-focused L2 instructional settings indicates that learners would like to be corrected more than their teachers think is good for them (Magilow, 1999; Schulz, 2001), which reveals a ‘miscalibration’ (Winne, 2004) between students’ and teachers’ beliefs about the instructional goals of CF. Such miscalibrations may be detrimental to CF effectiveness.

‘Usage of CF’, which we will define here broadly as what learners do with or in response to CF, comprises diverse constructs, namely 1) noticing of CF, 2) uptake and 3), in CALL settings, use of CF options. These constructs have been measured either through self-report instruments (such as stimulated recall and think-aloud protocols) or on the basis of behavioral data (including log files and eye-tracking data).

First, as for noticing, the research that taps into correction episodes in communicative interactions has gathered consistent evidence that feedback needs to be sufficiently explicit in order to be noticed. For instance, although the research on corrective recasts in L1 development has produced promising results and has spurred continued research in communicative L2 settings over the last decades (for a contrastive review see Nicholas, Lightbown, & Spada, 2001), L2 learners typically do not notice recasts if these lack perceptual salience (e.g., Lai & Zhao, 2006), if they are unsystematic (Nicholas et al., 2001) or non-contingent (i.e. if they do not immediately follow the erroneous response) (Lai, Fei, & Roots, 2008) or if they are long and involve many changes to the original utterance (Philp, 2003). Research on synchronous computer-mediated communication (CMC) has found that recasts which are textually enhanced are also associated with higher levels of awareness at the level of understanding than non-enhanced recasts.
In addition, it has been reported that learners have difficulty in identifying the linguistic focus of implicit CF types, especially for morpho-syntactic features of the target language (Mackey, Gass, & McDonough, 2000), and that even in the case of explicit recasts learners notice semantic and syntactic problems more easily than morphological ones (Smith, 2012).

A second construct that can be put under the umbrella of CF usage is uptake, defined as learner utterances in response to CF (Sheen, 2011, pp. 7–8). It is important to note that uptake is considered evidence of whether CF has been noticed, not whether it has facilitated development (Mackey & Philp, 1998). Uptake is known to be facilitated especially by CF types that are explicit and/or contain detailed information, which comprise techniques such as elicitation, corrective repetitions, metalinguistic feedback (Heift, 2004; Lyster & Ranta, 1997) and explicit recasts (Sheen, 2006). Next, recasts in general have proven more successful for eliciting uptake of lexical features than for grammatical features of the L2, although teachers use them widely for correcting grammatical errors (Mackey et al., 2000; Sheen, 2006). Heift (2001) describes metalinguistic feedback strategies in tutorial CALL activities for L2 grammar development, and concludes from the preponderance of learners’ repair movements (i.e., successful uptake) that learners attended to this feedback. In a similar research setting, Heift (2004) considered that uptake may be determined by two learner characteristics, viz. gender and language proficiency, but found no relation.

Third, CALL settings afford to investigate learners’ use of CF options, i.e., non-embedded support that comes with CF. Feedback sessions in CALL environments (particularly in tutorial CALL) may provide the learner with options, such as the possibility to see the location of the error, metalinguistic prompts, more extended grammar explanations and correct responses, which are all typically seen as comprising CF (Ellis et al., 2006). CALL research in this area has focused on the relation between usage of CF options and individual differences – this research focus is motivated by the desire to come to an understanding of what makes learners seek additional feedback. Heift (2006) showed that students’ usage of context-sensitive grammar help following CF is contingent upon the level of detail in CF and upon proficiency level: learners that were confronted with less detailed feedback and beginning learners tended to make more use of the error-specific help pages. Heift (2002) found that when introductory level university students of German were shown metalinguistic CF in tutorial grammar activities, the majority of students sought to correct errors mainly without relying on the use of correct answers. Heift (2002) found that when introductory level university students of German were shown metalinguistic CF in tutorial grammar activities, the majority of students sought to correct errors mainly without relying on the use of correct answers.

In addition, there appeared to be a relation between the learners’ strategies and their performance as measured by the system: learners that peeked at correct answers frequently, either in response to CF or without submitting in the first place, were low to mid performers; students that generally attempted to correct errors themselves and sometimes requested correct responses were mid-performers; and the students that virtually never relied on the usage of correct responses ranged from mid to high performance. Heift’s (2002) study corroborates previous findings by Brandl (1995), who concluded that students’ previous performance in class determined their usage of feedback options in tu-
DETERMINANTS FOR USAGE OF OPTIONAL FEEDBACK

Editorial grammar activities: low achievers looked up correct answers more often, whereas high-achieving students showed more willingness to engage in the correction process. Brandl also hypothesized that the low-achieving learners lacked adequate cognitive and motivational processing, and consequently he made a plea for more research into the relation between learners’ usage of CF and individual differences, in particular motivational variables. In the following section, we will review some of the more general literature on tool use in educational technology that might inspire CALL research in this area.

Tool Use in Computer-assisted Learning

As indicated in the introduction, the recent research on tool use in computer-assisted learning has begun to map, in an attempt to approach a detailed understanding of tool use, the complex relation between characteristics of the learning task, tool features, and learner characteristics. Learner characteristics that are thought to determine the usage of tools include prior knowledge, metacognitive skills and knowledge (including conceptions about the usefulness of instructional interventions), the functionality which learners attribute to specific tools in specific learning environments, and ‘motivation’ (Clarebout & Elen, 2006, 2009). The latter two constructs may both be considered perceptions, viz. perceptions about the usefulness of tools (perceived usefulness), and perceptions about the learner’s self, respectively. Perceptions are thought to emerge in the dynamic interaction between the learner and his or her environment, and are typically measured by means of self-report (questionnaires and/or interview data).

The construct of perceived usefulness originates in expectancy theory and is central in Davis’ technology acceptance model (TAM) (1989), which posits that users’ behavior (i.e. their use of technology) can be predicted by how useful they find the system and how easy they find it in actual usage, with perceived ease of use functioning as a causal antecedent of perceived usefulness. In the educational technology field, Lust, Elen, and Clarebout (2011) found evidence for the explanatory power of perceived usefulness with respect to students’ actual usage of webcasts in a blended learning course.

Next to perceptions about usefulness and ease of use of tools, learners’ perceptions of themselves might hint at their motivation for using (or not using) specific tools. This reasoning is reflected in current research on help-seeking (for a review see Aleven, Stahl, Schworm, Fischer, & Wallace, 2003), a line of educational research related to tool use which has started to investigate the relation of help-seeking strategies with learners’ achievement goal orientation. Achievement goals are typically bifurcated into mastery goals (also known as learning goals), which comprise ‘intrinsic’ goals focused on the development of competence or task mastery, and performance goals, which constitute a more extrinsic goal orientation, viz. demonstrating competence (e.g., relative to peers) rather than developing it (Elliot, 1999). Achievement goal theory assumes that mastery goals are associated with positive learning processes and outcomes (such as persisting through failure), whereas a performance orientation would lead to
less favorable behaviors and outcomes (e.g., lower effort in the face of failure, or surface processing of useful pedagogical materials). Along these lines, research on help-seeking has gathered evidence that mastery goals are typically associated with *instrumental help-seeking* (intended to promote learning, such as making use of hints), whereas performance orientation is more likely to be linked to *executive help-seeking* (intended to avoid work, such as peeking at correct responses) (Aleven et al., 2003). In addition, design features of learning environments may change learners’ achievement goal orientation, and subsequently such features could influence how individual learners seek help, e.g., by emphasizing performance or interpersonal performance comparisons (Karabenick, 2011). If learners’ use of optional CF can indeed be seen as related to instrumental help-seeking, then the literature on help-seeking may provide fertile theoretical models for individual difference research on the usage of optional CF in CALL, in particular as to how usage of elaborate feedback is driven by achievement goal orientation.

**Summary and Theoretical Framework**

To summarize, previous studies in the SLA literature have found that learners’ usage of CF is determined both by characteristics of CF and by individual differences. First, explicit and detailed feedback is more likely to facilitate noticing and uptake. Second, CALL research shows that students’ prior knowledge plays a role in how they engage with optional detailed feedback: beginning learners seem to request optional context-sensitive feedback more frequently than do advanced learners. And third, learners’ usage of correct response feedback has been shown to depend on their prior knowledge, or on their performance: low-achieving students look up correct responses more often than they work through feedback that does not give away the correct answer (i.e., output-prompting feedback). Thus, these learners may be considered to engage in executive help-seeking rather than in more independent problem-solving.

Hence, the outstanding question is why learners do or do not make use of certain feedback options. Of particular relevance is the case of weaker learners that resort to looking up correct responses, and hence make less frequent use of output-prompting feedback options in order to complete tasks. Potentially, these learners do not find output-prompting feedback useful, as they might lack the knowledge to cognitively process (meta-)linguistic explanations. Or, low-achieving students may lack “motivational processing” (Brandl, 1995, p. 207) to deal with such detailed feedback.

To address this question, the current study considers both perceived usefulness (of optional metalinguistic CF) and achievement goal orientation – as an operationalization of ‘motivation’ – as potential determinants of learners’ usage of optional metalinguistic CF, in addition to prior knowledge. The construct of perceived usefulness was chosen since it has been identified as a significant predictor of tool usage (Davis, 1989; Lust et al., 2011). We defined ‘motivation’ as achievement goal orientation (Elliot, 1999) because the learning environment which we wanted to evaluate contained features that stress achievement (see description in
DETERMINANTS FOR USAGE OF OPTIONAL FEEDBACK

next section) and may thus intensify differences between mastery-oriented learners and those that are primarily performance-oriented. Learners with the latter orientation are known to more frequently show executive help-seeking behavior (such as peeking at correct responses), whereas the former learners (oriented towards learning) would seem to take an interest in solving problems independently by attending to and working through detailed linguistic feedback.

Figure 1 summarizes the theoretical framework that forms the backbone of the current study. In addition to considering that the three key individual difference factors mentioned above, namely prior knowledge, perceived usefulness of CF and achievement goal orientation, may explain use of optional metalinguistic CF, we also need to recognize that the perceived ease of use of this CF might be determined by prior knowledge, more particularly explicit L2 knowledge. Carroll (1995) notes that CF is language about language and is thus “quintessentially metalinguistic in nature” (p. 76) – whether the CF includes metalinguistic information or not. This implies that learners need to be equipped with explicit L2 knowledge in order to decode feedback instances. Hence their explicit L2 knowledge might determine partly how easy it is for them to learn from CF, in addition to usability issues in interface design for example. In the following section, we will report on the design of a learning environment used in a pilot study aimed at exploring the relation between usage of CF options and prior (explicit) L2 knowledge, perceived usefulness of CF and achievement goal orientation.

**Figure 1. Theoretical framework**

**DESCRIPTION OF THE LEARNING ENVIRONMENT**

The learning environment used for the pilot study was a prototype of an online task-based tutorial CALL system for grammar practice in which learners played the role of a detective in ‘semi-open’ (Desmet, 2007) written activities, and had to solve a murder mystery by formulating responses that fell within the range of
predicted correct utterances. It utilized natural language processing (NLP) and crowdsourcing techniques to generate explicit embedded CF and non-embedded options that were deemed necessary to perform the tasks. The learning environment also contained features associated with gaming such as positive feedback. In this section, these features will be described in detail.

The learning environment was task-based (Ellis, 2003), as it was intended to capture learners’ interest by confronting them with a meaning-focused problem, which required them to work towards a non-linguistic outcome (i.e., solving the murder mystery through dialogue tasks), but on a lower level the activities involved writing responses to grammatical exercises integrated in the dialogues (hence also implying a strong focus on form). Although the unit of response was at the level of the utterance, which implies that many alternatives are possible, the range of appropriate utterances for particular exercises was constrained: first, by the immediately preceding and following utterances in the linear dialogues, which were provided by so-called non-player characters in the story (see Figure 2), and, second, by four grammatical topics in English, which are notoriously difficult for Dutch-speaking learners and for which errors are known to persist even in the speech of fairly advanced learners (Tops, Dekeyser, Devriendt, & Geukens, 2001). The grammatical topics and distribution of these topics in the exercises are shown in Table 1.

A string matching algorithm matched learners’ responses to a number of predicted responses for the grammatical exercises and for other ‘liaising interactions’ (i.e., sentences that did not target specific linguistic problems but that simply moved the dialogue forward). For the grammatical topics, both correct and incorrect utterances were predicted; the number of predicted responses for these interactions ranged between 1 and 19 (M = 5.5), depending on the scope of the exercises.

Table 1. Grammatical topics in the murder mystery

<table>
<thead>
<tr>
<th>grammatical topics</th>
<th>number of exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td>past time reference: simple past vs. present perfect</td>
<td>20</td>
</tr>
<tr>
<td>the quantifiers <em>some</em> and <em>any</em></td>
<td>5</td>
</tr>
<tr>
<td>modal verbs for ability, possibility, deduction</td>
<td>5</td>
</tr>
<tr>
<td>future tenses: <em>will</em> and <em>going to</em></td>
<td>1</td>
</tr>
<tr>
<td>( liaising interactions )</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>33</td>
</tr>
</tbody>
</table>
The string matching algorithm utilized simple string matching techniques (Levenshtein distance; see e.g., Lagatie & De Causmaecker, 2010) in combination with NLP (part-of-speech tagging and lemmatization) to compute the similarity between the learner’s response and each of the predicted responses. The outcome of the analysis included the closest match and the closest correct match, which were used to calculate a score and to generate immediate and explicit utterance-specific CF, as well as linguistic annotations at the level of the individual tokens in the learner’s utterance.

The utterance-specific CF comprised a visualization of the breakdown of the string matching procedure (see Figure 3), which was tested and improved in several iterations on the basis of experts’ comments. For each attempt the learner’s response was compared with the ‘closest match’: a predicted response that matched the learner’s response the closest. If the similarity was below a threshold of .5 (i.e., the learner’s response was less than 50 percent similar to the closest predicted utterance), the system showed an icon with a tooltip, which explained that the response had not been recognized and that the learner could try again (see first interaction in Figure 2). If the similarity was above .5 then the CF visualization routine underlined each of the tokens in the learner’s response that deviated from the corresponding token in the aligned version of the closest correct...
match. If the string-matching algorithm had not found a corresponding token in the aligned version of the closest correct match, then the CF visualization routine would simply show asterisks in that position. In a nutshell, the string matching algorithm detected words that were different from predicted words, words that were superfluous and words that were missing (Desmet, 2007); the CF visualization routine showed the location of the (potential) error in the sentence by means of underlining, a technique highly similar to highlighting CF in CALL and elicitation/repetition in classroom settings (Heift, 2004).

On top of the highlighting CF that was shown immediately after the learner’s response, non-embedded prompts were available at the level of the dissimilar tokens. These prompts were based on the part-of-speech and lemmatization analyses and included mainly metalinguistic terminology, which we thought could increase the chance that learners would be able to correct responses (e.g., “You might need a modal verb in this position.”), as well as encouraging and (arguably) humorous statements in line with the detective metaphor (e.g., “Oops, we expected nothing here. Try not to waste any words; detectives use as few words as possible.”).² Taking into account a pedagogical framework for the design of game-like activities for language learning (Purushotma, Thorne, & Wheatley, 2008), it was decided that this optional feedback would only be shown on the learner’s request and with varying degrees of specificity, dependent on the dissimilarity between the tokens (due to combinatorial differences with respect to lemma and part-of-speech). So, while feedback is generally embedded and hence not considered a tool (i.e., a non-embedded support device) (Clarebout & Elen, 2006), this additional feedback was available as a tool: learners had to click on the highlighted tokens in order to get access to context-specific detailed feedback.

In addition to the optional CF, other non-embedded help options were available. Based on usability testing, we assumed that the exercises would be quite challenging, especially the task of finding which words to use. The learners would be told they could find many words and chunks to be used in their responses ‘hidden’ in the utterances of the non-player characters, but still we decided to include two additional help options: access to the responses of others (represented as a red phone to call other detectives) – the learners seemed to like this idea – and an option to request a hint (represented as a key). The first help option is a crowdsourcing feature that showed the responses of other learners ranked by frequency (computed on the basis of the number of times the utterance had been submitted or chosen). When a peer response was chosen, it was evaluated in the same way as a free response, so learners would also get CF if this response was not correct. Each time the second option was clicked, learners received a correct word they could use in their response; and the option could be used until all words for a sentence were disclosed. Learners did not lose points for using any of these two options, and these were available both before and after each attempt.

Finally, the learning environment also contained features associated with gaming. First, the format of the learning environment is related to that of the interactive participatory drama, a genre which has a long history in CALL and which
has been discussed in game-based learning venues (Hubbard, 2002). In acknowledgment of the skills and effort required to write an engaging drama that would withstand the inclusion of grammar practice activities, the first author wrote a story on the basis of Edgar Allan Poe’s *The Murders in the Rue Morgue*. Secondly, the system computed scores for the learners’ individual responses, and represented these scores as “ideas” (light bulbs), adapted to the detective metaphor. Next, after each task/dialogue, learners would be given pieces of evidence on the murder mystery on the basis of how many “ideas” they had gathered. Moreover, accumulating points over time would result in increases in the learner’s level: they would start out as constable, then become inspector, etc., and eventually end up as superintendent. The level increase thresholds were balanced and play-tested in order to make sure that learners in our target audience would actually experience these level-ups while working through the tasks. Additionally, there was a leaderboard feature: learners could request the scores of the 10 highest-ranked peers by clicking on their own accumulated score. Finally, as is customary in game design, all of these instances of “positive feedback” were visualized rather excessively (Juul, 2010, p. 45) – episodes containing perfect responses and level-ups would include messages of verbal praise such as “Bravo!”, and were animated using the jQuery library for HTML5-compliant websites. Figure 3 summarizes the design features of the learning environment.

Figure 3. Design features of the learning environment (features in red, e.g., “request optional CF”, are non-embedded)
PILOT STUDY

Aim & Research Questions

The pilot study had a few practical aims and was also intended to empirically explore the theoretical background on CF usage outlined above. As to the practical aims, first, we wanted to evaluate whether students found the metalinguistic prompts at all usable. Secondly, we wanted to check whether the self-report instruments (see below) were reliable for use in future experiments. A third aim was to collect typical responses from learners to expand the domain model of the tutoring system, i.e., to populate the content database on the basis of learner language with additional and more evidence-based instances of correct and incorrect responses, and to evaluate the accuracy of the string matching algorithm (an investigation which is beyond the scope of this paper). Another objective was to evaluate the gaming features (positive feedback specifically), in consideration of providing learners with different positive feedback types (or none at all) in future experiments. As a final practical aim, we wanted to evaluate the technology in the setting of a typical secondary school classroom in order to detect potential performance problems with the software.

Additionally, the pilot study aimed to explore learners’ usage of optional CF in relation with its perceived usefulness, with prior knowledge and achievement goal orientation (see the problem statement above in the section ‘Background Research’). We also hypothesized that the perceived usefulness of the CF may be determined by the perceived difficulty of the task, taking into account the challenge for learners to construct responses that fell in the scope of predicted utterances in these semi-open activities; we predicted that the CF would not be found useful if the task was too difficult. Next, we also considered that frequent executive help-seeking strategies (requesting hints and peer responses) could be associated with a performance goal orientation. A final aim was to empirically explore the relation between usage of optional CF, use of hints, use of peer responses and attempts per exercise. The research questions were as follows:

1. How useful do learners find the CF? How is the perceived usefulness of CF related to its perceived ease of use, and to the perceived difficulty of the task?
2. How is perceived ease of use of CF related to prior explicit L2 knowledge?
3. How frequently do learners use the optional CF? How is this usage related to the perceived usefulness of CF, to prior explicit L2 knowledge and to achievement goal orientation?
4. How is the usage of hints and peer responses related to achievement goal orientation?
5. What is the relation between usage of the optional CF, use of hints, use of peer responses and attempts per exercise?
Participants, Procedure, and Instruments

The study was carried out in May 2012 in two classes in a secondary school in Kortrijk, Belgium. Thirty-six Dutch-speaking high-intermediate learners of English in the 5th and 6th form of the Modern Languages program were invited to the study through their teachers. There were 29 girls and 7 boys, and although we did not have their exact ages, these would typically be in the 16-18 year range.

One week before the learners worked with the online learning environment, they took an English grammar test that was intended to measure their prior explicit L2 knowledge. The test included adapted versions of the metalinguistic knowledge test (MKT) and grammaticality judgment test (GJT) published in Ellis (2009). These tests cover a wide range of grammatical structures (17 in total) that are known to be universally problematic to learners of English as L2 at various stages in their development and may, hence, provide a representative performance measure of linguistic ability. In the MKT, participants are presented with ungrammatical utterances for these structures and are required to select for each utterance the rule that best explains the error out of a list of four options. As the MKT draws heavily on learners’ knowledge of metalanguage, involves a high degree of awareness, and focuses attention on form, it is considered to be a measurement of learners’ explicit L2 knowledge. The grammatical structures that form part of the MKT are also included in the GJT. In this task, participants decide whether utterances are well-formed or deviant. The GJT may measure implicit or explicit L2 knowledge, depending on the conditions of the task, more particularly the time learners are given to make the judgments, and the nature of the stimuli. If participants are given only limited time to respond, they may need to rely primarily on their implicit (intuitive) L2 knowledge. Conversely, if they have unlimited time to judge utterances, especially ungrammatical ones, it is more likely that they are relying on explicit L2 knowledge (Loewen, 2009). Ellis (2009) found very strong significant correlations between the MKT and the ungrammatical items on the GJT if the latter was untimed, which suggests that learners’ performance on the ungrammatical items of the untimed GJT also provides a measure of their explicit L2 knowledge.

For this study, we selected the MKT and the untimed version of the GJT, as we presumed that learners’ explicit L2 knowledge might affect how they would use optional metalinguistic CF. Also, these tests were chosen, rather than other L2 knowledge tests, as they had been found highly reliable in a previous study (Ellis, 2009). For time constraints, the GJT was reduced to 34 utterances (half the size of the original test), but covered the 17 grammatical topics of the original test, and we also left out items that measured participants’ self-reported use of rules and how certain they felt about their responses. For the MKT, we used only the first part, which consists of 17 multiple-choice items that target knowledge of rules and which requires understanding of metalinguistic terminology. For both tests, the items and terminology were slightly altered to more accurately reflect how the grammar had been taught in class; these adjustments were done on the basis of comments from one of the participating teachers.
One week later, the learners worked in the online environment in a computer room at school. The first author guided learners through the environment by means of a slideshow with screenshots of its key features, emphasizing the feedback and other help options, and briefly reviewed the grammar rules of the four topics that the learners were about to practice. The final slide contained a summary of this walkthrough including the suggestion to make use of the feedback and other help options, which remained visible while learners were working through the tasks. All instructions were given in English. The researcher and the teachers helped the learners if they experienced technical difficulties, but did not intervene for problems with grammar.

The participants were not familiar with the murder mystery, and it was the first time they worked in the environment. Four dialogue tasks were available in the story, through which the learners would navigate in a fixed order (each subsequent task would only become available after the learners had finished the previous one). Each dialogue task contained between 7 and 9 exercises. The number of attempts for each exercise was not restricted; the system would only move on to the next exercise automatically if the score was perfect. The metalinguistic terminology in the non-embedded CF prompts (e.g., “infinitival to”, “linking word”, “modal verb”) had been adapted to the terminology used in the learners’ grammatical compendium in their course book.

In line with recommendations from Fischer (2007) to put the study of learners’ reliance on help features in CALL on more solid empirical footing, all learners’ actions (except keystrokes) were logged. Most learners used a personal online account (Facebook or Gmail) to log into the system; three learners who did not have an online account or who had difficulties logging in used one of the spare accounts (Gmail) provided by the researcher. The first class practiced the grammar problems in the murder mystery for about 40 minutes; due to time constraints the second class only practiced 25 minutes. Because not all learners completed the four tasks, large differences were expected between learners for the number of completed tasks, completed turns and number of attempts. This needed to be taken into account for the analyses (see next section on ‘Data Preparation and Preliminary Analyses’).

After having worked in the online environment, the learners filled out a questionnaire (in their L1) that contained four sections with 7-point Likert-scale items. The first section targeted learners’ achievement goal orientation while working in the environment and was based on and translated into Dutch from the 3x2 achievement goal model (Elliot, Murayama, & Pekrun, 2011). It consisted of items on 6 goal constructs: task-approach and task-avoidance (i.e., goals focused on learning), self-approach and self-avoidance (i.e., performance goals focused on improving previous performance) and other-approach and other-avoidance (i.e., performance goals focused on outperforming others). The second section of the questionnaire was based on the TAM (1989) and comprised items on the perceived ease of use of the optional metalinguistic CF as well as items on its perceived usefulness. For each construct we included multiple items, in order to provide reliable and valid measures for each of the constructs involved (Dörnyei,
To further improve reliability, some items in the TAM scale were negatively worded in order to be able to detect participants that would consistently pick, e.g., a 6 or 2 without reading the items. Table 2 shows an overview of the constructs, and for each construct the number of items included in the questionnaire as well as an example item. In the final two sections, learners filled in additional single 7-point Likert-scale items on how motivating they had found particular features in the environment (the murder mystery, the collection of evidence, the accumulated score, leveling up, the leaderboard, collaborating with others, the corrective feedback), and how difficult they had found the task (in general, with respect to finding the right words, with respect to the grammatical problems; one item for each construct).

Table 2. Overview of constructs included in questionnaire

<table>
<thead>
<tr>
<th>scale</th>
<th>construct</th>
<th>number of items</th>
<th>example item</th>
</tr>
</thead>
<tbody>
<tr>
<td>3*2</td>
<td>achievement</td>
<td>3</td>
<td>I found it important to get a lot of questions right.</td>
</tr>
<tr>
<td></td>
<td>goal model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>task approach</td>
<td>3</td>
<td>I tried to avoid getting a lot of questions wrong.</td>
</tr>
<tr>
<td></td>
<td>task avoidance</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>self approach</td>
<td>3</td>
<td>I tried to do well relative to how well I did on previous attempts and exercises.</td>
</tr>
<tr>
<td></td>
<td>self avoidance</td>
<td>3</td>
<td>I tried to avoid performing worse in comparison with previous attempts and exercises.</td>
</tr>
<tr>
<td></td>
<td>other approach</td>
<td>3</td>
<td>I tried to do better than my peers.</td>
</tr>
<tr>
<td></td>
<td>other avoidance</td>
<td>3</td>
<td>I tried to avoid doing worse than my peers.</td>
</tr>
<tr>
<td>technology acceptance</td>
<td>perceived ease of use</td>
<td>6</td>
<td>I found it difficult to understand the grammatical feedback.</td>
</tr>
<tr>
<td>model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>perceived usefulness</td>
<td>6</td>
<td>The grammatical feedback helped me to learn why I had been wrong.</td>
</tr>
</tbody>
</table>

RESULTS

Data Preparation and Preliminary Analyses

We tested the reliability of the grammar tests using the Kuder Richardson’s Formula 20 (KR-20), which calculates the homogeneity of a test with dichotomous measures (Kuder & Richardson, 1937). Tests with a KR-20 coefficient over .90 are considered homogeneous and hence reliable. Analysis revealed that the grammar tests did not reliably measure learners’ explicit L2 knowledge (KR-20 $\alpha = .02$ for the ungrammatical items in the GJT; KR-20 $\alpha = .04$ for the MKT). Inspection of the sum scores for the tests indicated that overall, the learners’ scores on the tests were rather high (for the ungrammatical items in the GJT: $M = .91, SD = .06$; for the MKT: $M = .74, SD = .10$). Possibly the low reliability of the grammar tests was due to the fact that there was little variation between the learners
in terms of their explicit L2 knowledge. Yet, we decided to keep the tests in our analyses for two reasons: they had been carefully constructed and found reliable in previous studies (Ellis, 2009), and they had been adapted to the learners’ curriculum with the help of one of the teachers, so they would reflect how grammar was typically taught and assessed in class.

The self-report measures (questionnaires), on the other hand, were highly reliable. For the questionnaires, we computed Cronbach’s $\alpha$, which is a measure of the internal consistency of a scale or any of its subscales/constructs (Cronbach, 1951). Scales with $\alpha$ coefficients above .70 are considered reliable, whereas scales with $\alpha$ coefficients below .60 should be treated with caution (Dörnyei, 2003). For achievement goal orientation Cronbach’s $\alpha$ was between .78 and .96 for the items on the subscales; for the TAM Cronbach’s $\alpha$ was .95 for the items on perceived usefulness and .84 for the items on perceived ease of use. Hence, we considered the subscales for the constructs related to learners’ perceptions reliable. The subscales were created by taking the means of the corresponding items.

Inspection of the correlation matrix for the subscales of the achievement goal orientation questionnaire (see Table 3) revealed significant positive intercorrelations between many of the subscales, particularly between task- and self-focused goals, which suggest that the latter goals “emerge from very similar dispositions in general” (Elliot et al., 2011, p. 641). Other-approach goals were unrelated to the former goals; this may have emerged from the fact that the learning environment stressed interpersonal comparisons through the leaderboard. Hence, for further analysis, we chose two approach goals that were unrelated to each other: task-approach, focused on learning, and other-approach, focused on outperforming others.

Table 3. Pearson’s correlation coefficients for achievement goal motivation, adjusted for multiple comparisons using Holm’s Method (***$p < .01$; *$p < .05$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. task approach</td>
<td>—</td>
<td>.84**</td>
<td>.76**</td>
<td>.74**</td>
<td>.28</td>
<td>.46*</td>
</tr>
<tr>
<td>2. task avoidance</td>
<td>—</td>
<td>.74**</td>
<td>.80**</td>
<td>.32</td>
<td>.57**</td>
<td></td>
</tr>
<tr>
<td>3. self approach</td>
<td>—</td>
<td>.88**</td>
<td>.25</td>
<td>.46*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. self avoidance</td>
<td>—</td>
<td>.32</td>
<td>.48*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. other approach</td>
<td>—</td>
<td>.84**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. other avoidance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As to the log data, as was expected, there were large differences between learners with respect to the amount of completed tasks, completed turns and attempts, which reflects differences in the amount of time which the learners had spent in the online environment. So, after extraction, the log data were normalized in order to even out these differences. All learners consulted the optional CF at least once.
Usage of optional CF was calculated by dividing the number of times learners had clicked a token to see metalinguistic prompts by the amount of highlighted tokens for which a metalinguistic prompt had been available. Next, learners’ usage of hints and their usage of others’ responses were computed by dividing the number of times they had requested these help options by their individual total number of attempts on the level of the exercises.

**Findings**

As for research question one, the results indicate that the majority of learners found the CF quite useful, with a median of 4.75 ($M = 4.51$ ; $SD = 1.12$), scores ranging between 2 and 6.67, and 69% of the participants scoring 4 or higher on the perceived usefulness scale. Further, the correlation matrix in Table 4 shows a significant positive correlation between perceived usefulness of CF and perceived ease of use of CF, and a significant negative correlation between perceived lexical difficulty and perceived ease of use of CF ($r = -.42$ ; $p = .05$). The latter finding indicates that learners who experienced less difficulty in getting the words right in the dialogue exercises found the CF more easy to use, and vice versa.

Table 4. Pearson’s correlation coefficients for research question 1, adjusted for multiple comparisons using Holm’s Method (**$p \leq .01$; *$p \leq .05$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. perceived usefulness</td>
<td>—</td>
<td>.62**</td>
<td>-.40</td>
<td>-.11</td>
</tr>
<tr>
<td>2. perceived ease of use</td>
<td>—</td>
<td>—</td>
<td>-.42*</td>
<td>-.31</td>
</tr>
<tr>
<td>3. perceived lexical difficulty</td>
<td>—</td>
<td>—</td>
<td>.45*</td>
<td>—</td>
</tr>
<tr>
<td>4. perceived grammatical difficulty</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The second research question focused on the relation between perceived ease of use of CF and prior explicit L2 knowledge as measured by the MKT and by the ungrammatical items on the GJT. Table 5 shows that there were no significant correlations between these variables.

Table 5. Pearson’s correlation coefficients for research question 2, adjusted for multiple comparisons using Holm’s Method (**$p \leq .01$; *$p \leq .05$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. perceived ease of use</td>
<td>—</td>
<td>.03</td>
<td>-.10</td>
</tr>
<tr>
<td>2. GJT (ungrammatical items)</td>
<td>—</td>
<td>—</td>
<td>.07</td>
</tr>
<tr>
<td>3. MKT</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
For the third research question on usage of optional CF, descriptive statistics show high variability between learners in terms of their usage of CF ($M = .40; SD = .26$; range between .03 and 1.02). The correlation matrix for the third research question (see Table 6) shows that usage of optional CF is only – but strongly and significantly – related to explicit L2 knowledge as measured by the MKT. This signifies that learners with higher explicit L2 knowledge used the optional metalinguistic CF more often.

Table 6. Pearson’s correlation coefficients for research question 3, adjusted for multiple comparisons using Holm’s Method ($**p \leq .01; *p \leq .05$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. usage of optional CF</td>
<td>—</td>
<td>-.08</td>
<td>.59**</td>
<td>.03</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>2. perceived usefulness</td>
<td>—</td>
<td>-.05</td>
<td>-.23</td>
<td>.44</td>
<td>-.16</td>
<td></td>
</tr>
<tr>
<td>3. MKT</td>
<td>—</td>
<td>.09</td>
<td>.15</td>
<td>.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. GJT (ungrammatical items)</td>
<td>—</td>
<td>-.13</td>
<td>.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. task approach</td>
<td>—</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. other approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For research question 4, no significant relations were found (see Table 7).

Table 7: Pearson’s correlation coefficients for research question 4, adjusted for multiple comparisons using Holm’s Method ($**p \leq .01; *p \leq .05$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. usage of hints</td>
<td>—</td>
<td>.16</td>
<td>-.12</td>
<td>-.21</td>
</tr>
<tr>
<td>2. usage of responses from peers</td>
<td>—</td>
<td>.00</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>3. task approach</td>
<td>—</td>
<td></td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>4. other approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results for research question 5 reveal significant and medium-sized negative correlations between usage of CF and usage of hints ($r = -.38; p = .05$) and between usage of CF and usage of responses from peers ($r = -.39; p = .05$) (see Table 8). This suggests that students who requested the additional CF more often used both hints and responses of others less often, and vice versa. In addition, there are significant correlations between the number of attempts per exercise and usage of help options: the relation goes in the negative direction for usage of hints and usage of responses from peers; the relation is positive for usage of CF. Thus, we may derive that learners who made more frequent use of the hint and peer response options were engaged to a lesser extent in the correction process,
whereas learners who used the optional metalinguistic CF more often showed more willingness to correct their responses autonomously.

Table 8: Pearson’s correlation coefficients for research question 5, adjusted for multiple comparisons using Holm’s Method (**p ≤ .01 ; *p ≤ .05)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. usage of CF</td>
<td>—</td>
<td>-.38*</td>
<td>-.39*</td>
<td>.49**</td>
</tr>
<tr>
<td>2. usage of hints</td>
<td>—</td>
<td>—</td>
<td>.14</td>
<td>-.46*</td>
</tr>
<tr>
<td>3. usage of responses from peers</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-.58**</td>
</tr>
<tr>
<td>4. attempts per exercise</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

**DISCUSSION AND SUGGESTIONS FOR FUTURE RESEARCH**

This pilot study was aimed at identifying individual difference factors that might explain learners’ usage of optional metalinguistic CF. First, we found that usage of this CF was related to prior explicit L2 knowledge as measured by the metalinguistic knowledge test described in Ellis (2009): more ‘advanced’ learners made more frequent use of the optional metalinguistic CF. Given the low reliability of the grammar test in this particular context, this finding needs to be taken with care. Nonetheless, it seems to be consistent with previous research in CALL which has found that ‘high-achieving’ learners make more use of output-prompting CF (Brandl, 1995; Heift, 2002), but contradicts Heift’s (2006) findings: there, language proficiency was inversely related to usage of grammar help following feedback. Theoretically, one might expect the latter, viz. that more advanced learners would need less feedback than less advanced learners in order to successfully complete a specific task – this prediction has been articulated most clearly by sociocultural approaches to feedback usage (e.g., Aljaafreh & Lantolf, 1994). Hence, our finding is somewhat hard to interpret in the face of current research and theory, but it may be explained by our specific operationalization of language proficiency as explicit L2 knowledge: learners equipped with explicit L2 knowledge might have been more able to decode the optional metalinguistic feedback and, hence, used it more. Future research, however, could include other language tests as measures of prior knowledge, in order to yield a more complete picture of the learners’ proficiency levels.

On the basis of our findings, we can neither confirm nor deny that usage of optional CF depends on its perceived usefulness, a prediction made by the TAM. The lack of relation between perceived usefulness and actual usage is in line with previous findings in CALL, and can be explained by the hypothesis that learners’ perceptions may be inaccurate and that “researchers should be ever mindful of the discrepancy between statements of learners’ perceptions/beliefs and their actual behaviors” (Fischer, 2007). This explanation, however, makes it difficult to put the widely supported TAM to the empirical test in educational settings. In this particular case, another explanation might be in order: learners might have found
the embedded CF (highlighting) sufficiently useful, so that they did not need to use the non-embedded CF. This explanation is supported by the observation that in general, the participants were quite advanced learners of English. For these learners, possibly, the optional CF did not add much information to the embedded CF, save for a hint on which part of speech was expected for a particular position in the utterance. The optional CF did not include extended grammar rules – in fact, grammar rules were orally presented to the learners before practice. So, the embedded CF might have served as a proxy for the grammar rules (as we had intended), which could have constituted sufficient learning support for this particular task and for these particular learners. Subsequently, when responding to the items on perceived usefulness in the post-questionnaire, the participants might have restricted their judgments to the highlighting CF, even though the questionnaire items were introduced by a screenshot showing a highlighted token with the optional CF. Hence, this would have yielded an inaccurate measurement of perceived usefulness of optional metalinguistic CF. This does not mean that future research should abandon investigating learners’ perceptions, but that improved versions of the questionnaire should more clearly distinguish between the embedded and non-embedded CF. Also, additional instruments could be used, such as post-experimental interviews to gain a more detailed understanding of learners’ perceptions, or eye-tracking to measure another aspect of CF usage, namely noticing.

Similarly to perceived usefulness, learners’ achievement goal orientation could not explain their use of optional CF, nor could it explain usage of hints and peer responses. In other words, there was no evidence that task-approach was associated with instrumental help-seeking behavior (intended to promote learning), which we hypothesized would be reflected in the usage of optional CF, and no evidence that executive help-seeking (in this case frequent use of hints and peer responses) was associated with a performance orientation (other-approach) (see also Baker et al. (2008) for the lack of association between these variables). We did, however, find that both use of hints and use of peer responses were inversely related to usage of optional CF and to the number of attempts per exercise, which could support the distinction between executive and instrumental help-seeking.

On a more general plane, the lack of relations between learners’ perceptions (perceived usefulness and achievement goal orientation) and optional CF usage might be attributed to the limited sample of participants and short period of practice characteristic of this pilot study. Learners may not yet have been sufficiently familiar with the features of the learning environment. Their perceptions might not have stabilized and, therefore, may have been inaccurate. The following observation illustrates this hypothesis. During the practice sessions one learner remarked that he refrained from using the hint option because he thought he would cheat and lose points. This was not the case – which implies that this learner’s perceptions were not ‘calibrated’ (Winne, 2004) to the designed goals of the environment – but it constitutes a plausible belief in the context of gaming ecologies. Thus, in future studies, participants might need to be given more technical guidance or additional time to practice in the environment, before their percep-
tions are measured. Lastly, future studies could consider exploiting the gaming features inherent in the environment in an attempt to incentivize learners’ usage of optional (metalinguistic) CF. Performance-oriented learners might use optional help features more often if they get rewarded for correct responses and if the help offered is actually useful for solving problems. In order to empirically investigate the effects of game-like features on learner behavior, positive feedback and competition mechanisms could be implemented in separate experimental conditions.

ACKNOWLEDGEMENTS

The development of the learning environment used in this study was partly funded by the iMinds LLINGO project, co-funded by iMinds (Interdisciplinary Institute for Technology), a research institute founded by the Flemish Government.

NOTE

1 Note that the notion of tool in this literature is different from Levy’s (1997) conceptualization of the term in the field of CALL, and is more closely related to the tool and monitor functionality types of CALL applications described in Colpaert (2004).

2 For this feedback message, we acknowledge our indebtedness to a similar design feature in Sanders’ and Sanders’ (1995) classic tutorial CALL system SPION.

REFERENCES


DETERMINANTS FOR USAGE OF OPTIONAL FEEDBACK


