Sabrina Wilske, Magdalena Wolska

Meaning versus Form in Computer-assisted Task-based Language Learning: A Case Study on the German Dative

We report on a study which investigated the effects of three types of feedback realized in instructional dialogues with a computer-based language learning system for German. The interaction was framed within a directions giving task and the linguistic form in focus was the dative case in prepositional phrases. The feedback types differed with respect to the focus they put on form versus meaning and the explicitness of feedback in response to learner errors. The results of the study suggest that a stronger focus on form is related to greater accuracy gains in using the form. The integration of incidental focus on form within a primarily meaning-based task increases accuracy as well, however to a lesser extent.

1 Motivation and research questions

One of the research objectives actively investigated in the second language acquisition (SLA) community is to determine what types of instruction are most effective for foreign language learning. Generally speaking, language instruction can give priority to formal aspects of the language or to meaning and content. Long (1991) proposed a distinction between three types of instruction in terms of emphasis on form versus meaning: Instruction may require the learner to focus on meaning, on form, or both at the same time. While Focus on Meaning (FonM) does not draw learners' attention to linguistic forms at all, Focus on Forms instruction (FonFs) focuses on forms in isolation, providing no or only limited meaningful context. Focus on Form instruction (FonF) tries to integrate meaning and form by drawing learners' attention to linguistic forms as they arise within primarily meaning-oriented interaction.

Focus on Form is often realized within communicative interactions which provide opportunities for the learner to produce comprehensible output as well as to modify their output in response to feedback, thereby stimulating learning (Long, 1981; Krashen, 1985; Swain, 1985). The communicative approach advocates the use of goal-oriented communicative activities, *tasks*, in foreign language learning (Long, 1991; Ellis, 2003). Communicative goals of tasks should be framed in real world situations which elicit the use of the developing language from the learners. Important definitional properties of tasks are (1) primary focus on meaning, (2) clearly defined communicative outcome, and (3) free use of linguistic forms which the learner chooses. The third point gives rise to a potential problem when, as part of a pedagogical strategy, a specific grammatical structure of a language is targeted: Because learners are free to use any forms they want, one cannot guarantee that they will use the forms of interest. Therefore *focused tasks* have been proposed as an attempt to integrate forms and meaning. Focused tasks are designed in such way that learners are likely to use a specific target structure thereby improving its mastery.

One of the factors contributing to the effectiveness of communicative interaction is the type of feedback learners get in response to non-target-like contributions: (1) explicit vs. implicit feedback, and (2) prompting for a correction or not. Instruction is considered explicit if it contains an explanation of the language phenomenon in question or asks learners to attend to particular forms in the target language. It is considered implicit otherwise (Norris and Ortega, 2001). Corrective reformulations of learner's utterances or their parts, so called *recasts*, provide implicit feedback without prompting for correction and thus do not disturb the task-level conversation. By contrast, *metalinguistic feedback* (comments or questions related to the error which do not explicitly provide the correct form) is explicit and thus temporarily shifts attention from meaning to form. While both of these feedback types have been previously investigated in the classroom context (see, for instance, (Lyster and Ranta, 1997)) there has been little research into the efficacy of different feedback types in computer-based dialogic language instruction. A previous study by Norris and Ortega (2001) comparing the efficacy of different types of instruction (primarily non-computer based) suggests a slight advantage of explicit instruction over implicit by showing that the former results in higher test scores. The same study suggests that FonF and FonFs have equivalent effects. Ferreira (2006) found that when a computer interface is involved, feedback which prompted learners to correct their error yielded more learning gains than feedback which provided the correct target form.

In this paper we report on a study which compared the effects of three types of computer-based dialogue activities which differ in terms of the degree of focus on form vs. meaning, the degree of explicitness of feedback, and correction prompting strategies, on the acquisition of foreign language structures. The activities were performed using a type-written computer-based dialogue system. The interaction with the system was framed within a directions giving task and the linguistic form in focus is the German dative case in prepositional phrases. Our research questions were:

(1) Does computer-based task-oriented interactive instruction help learners of German improve accuracy on the use of the dative case in prepositional phrases?

(2) Is there a difference in the effect of free (FonF) vs. constrained (FonFs) type-written production on the acquisition of the dative in prepositional phrases?

(3) Is there a difference in the effect of implicit feedback (recasts) vs. explicit (metalinguistic) feedback on the acquisition of the dative in prepositional phrases?

The pedagogical goal was two-fold: Learners should improve their communicative skills in the scenario and their control of the target structure. In this paper we report the results on the latter.

In general, the idea of computer-based dialogues for foreign language learning is not new. Computer assisted language learning (CALL) has been an active research area for many years. With the progress in language technology the number of intelligent CALL systems, allowing learners to use natural dialogue, has been growing; see, for instance, (Holland et al., 1998; Harless et al., 1999; Seneff et al., 2004; Johnson et al., 2004). However, most systems are not built with the goal of transferring findings or testing hypotheses from the field of SLA; see (Petersen, 2010) for one exception.

Gender	Nominative	Dative PP	Translation
Masc	**der** Laden	hinter **dem** Laden	behind the shop
Fem	**die** Mensa	hinter **der** Mensa	behind the canteen
Neut	**das** Cafe	hinter **dem** Cafe	behind the cafe

Table 1: German dative in a prepositional phrase

Outline This paper is organized as follows: Section 2 describes the scenario, the target structure, and the types of instruction we evaluated. In Section 3 the implemented dialogue system is briefly outlined. Section 4 presents the design of an experiment we conducted. Section 5 summarizes the results of the study. In Section 6 we discuss the findings and conclude.

2 The Approach

In line with the focused tasks method, for the communicative instruction we selected a grammatical form and a task such that the form is natural to use within the task scenario and such that the scenario is meaningful and useful for the learner. We introduce the form and the task below.

2.1 The target forms and the tasks

The form: Dative case in prepositional phrases Among other uses, the dative case in German is required as an object of certain spatial prepositions. The dative case in German is marked morphologically on the gender-specific determiner of a noun phrase as well as on adjectives and in specific cases on the head noun. Table 1 shows the nominative and dative case forms (emphasized in bold) for the three German genders. Most locative prepositions used for describing static spatial relations require dative, among others, *vor* ('in front of'), *hinter* ('behind'), *neben* ('next to'), or *zwischen* ('between') The directional prepositions *zu* and *bis zu* ('to, towards') also require dative. These prepositions can be elicited in a task involving spatial descriptions.

The task: Giving directions We designed the directions giving task in a way so that it most efficiently attempts to elicit the forms of interest. The learner is presented with a simplified map of a fictitious campus, with buildings, other landmarks and a route to describe. Figure 1 shows the actual material used in our study. The scenario described when presenting the task is that the learner was stopped on the campus and asked for directions. The instructions explicitly request that the map provided be used and that the indicated route be described. The task description does not include any hints as to using prepositional phrases or paying attention to the dative case. The landmarks we used are balanced as to their gender and the gender is provided on the map. The

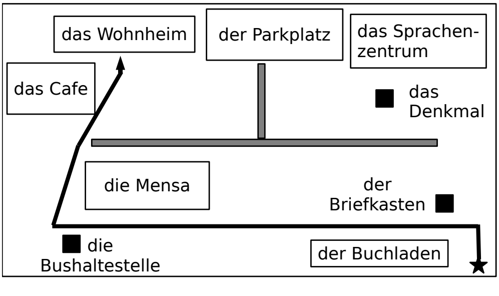

Figure 1: The map used in the "Directions giving" task

Dialogue strategy

```
interpretation-found = false
get-user-input

if user-input-parsed
   interpretation-found = true
else
   if keyword-match-found
      interpretation-found = true

if interpretation-found == true
   if TF-realized
      generate-feedback-recast
         or generate-feedback-metaling
   else
      elicit-TF
else
   output
     `Sorry, I didn't understand.'
```

Feedback functions

```
generate-feedback-recast:
 if TF-incorrect
   recast-TF
 prompt-for-next-contribution

generate-feedback-metaling:
 if TF-incorrect
   if first-trial
      output
        `<DET> in <NP> is not correct.'
   else
      output
        `<DET> is not correct either.'
        `Use dative!
   prompt-for-correction
 else
   prompt-for-next-contribution
```

Figure 2: Dialogue strategy in the two free production activities as pseudo-code.

route includes two points of direction change each at two landmarks and the target is placed close to two other landmarks. This setup makes it likely that the dative will be used when referring to a turn at a landmark in order to make the directions more precise and it also gives an opportunity to ask clarification questions when the learner does not supply the target forms. The two landmarks at each point of direction change are either both feminine or masculine, while all the landmarks close to the target have neuter gender. The learner has thus an opportunity to use the dative with all the genders, but in case they do not use locative prepositions we can also explicitly elicit all of these forms.

2.2 Task-based instruction

We designed and implemented three variants of communicative instruction: All variants involved a type-written dialogue with the system we built in order to perform the "Giving directions" task. The system controlled the interaction by means of a state-based dialogue model. The three variants of the instruction differed in the extent of freedom of language production they offered and the realization of form-focused feedback. In two variants of FonF activities the learners were able to freely formulate their dialogue contributions, *free production*, while in the third variant, *constrained production*, implementing FonFs, the learners' production was limited to supplying the target form (filling a gap) . The feedback in the latter activity variant was explicitly stating whether the supplied form was correct or not. The two free production variants differed with respect to the feedback they provided in response to incorrect forms: One implicitly corrected the error while maintaining the focus on the task-level conversation whereas the other explicitly pointed the learner to the error and demanded a correction thereby briefly focusing on the form. We elaborate on the properties of the respective system variants below.

Free language production–FonF In the free-production FonF activities the learners were able to type their utterances freely without any restrictions on the language used. The system implemented two input interpretation strategies: one based on a grammar with mal-rules and the other a fall-back strategy, based on fuzzy keyword matching (see Section 3). The system classified the learner's input into one of three categories ("TF" stands for "target form"): TF-realized-correct, TF-realized-incorrect, TF-not-realized. The high-level dialogue and feedback strategy of the system is summarized as pseudo-code in Figure 2.[1] If the learner's input was classified as not realizing the target form, the system tried to elicit it once by asking a clarification request, as exemplified in (1):[2]

(1)	**L:**	und dann nach links	*and then left*
	S:	[wo soll ich links?]$_{ELICIT}$	*where do I turn left?*

In case of learner errors in the target form (the TF-realized-incorrect category) the recast system provided implicit feedback by reformulating, *recasting*, the learner's utterance or parts thereof. Recasts were realized in a way so as to give them an appearance of implicit confirmation type of grounding moves, as in (2).[3]

(2)	**L:**	Gehen Sie hinter **das** Cafe nach links.	*Turn left, past the coffee-shop*
	S:	Okay, [hinter **dem** Cafe nach links,]$_{RECAST}$	*Okay, left past the coffee-shop*
		[und dann?]$_{PROMPT}$	*and then?*

The metalinguistic feedback system would explicitly state that there is an error, point to the location of the error and elicit a correction by the learner, as shown in (3). In case the learner should not succeed in correcting the error, the system would give a further hint, as in (4); cf. Figure 2.

(3)	**L:**	Gehen Sie hinter **das** Cafe nach links.	*Turn left past the coffee-shop*
	S:	['das' in 'das Cafe' ist nicht richtig.]$_{METALING}$	*'das' in 'das Cafe' is not correct.*
		[Bitte noch einmal!]$_{PROMPT}$	*Please try again!*
(4)	**L:**	hinter **den** Cafe nach links.	*left past the coffee-shop*
	S:	['den' in 'den Cafe' ist auch nicht richtig.]$_{METALING}$	*'den' in 'den Cafe' is not correct either.*
		[Nimm Dativ!]$_{PROMPT}$	*Use the dative!*

The system did not attempt to diagnose nor correct any other incorrect structures except those in focus. We anticipated that some learners might give a complete route description in one turn at the start of the dialogue. In order to ensure longer engagement, the system prolonged the interaction

[1] We omitted some system turns signaling non-understanding due to unknown words in order to simplify the presentation.

[2] **S** and **L** mark system and learner turns respectively.

[3] The bold emphasis did not appear in system output and is used here only to indicate the incorrect form and its correction via recast.

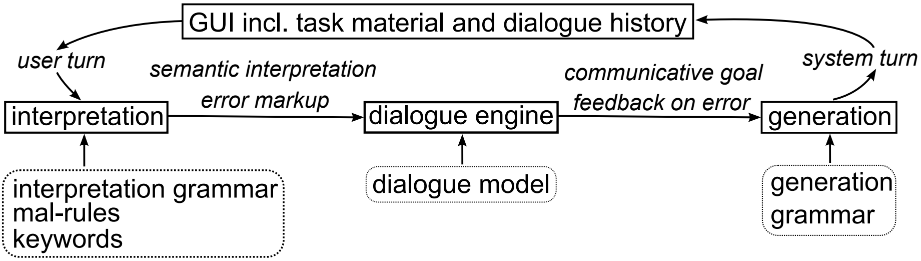

Figure 3: The system architecture; rectangles: modules, bottom part: resources, arrows: information flow

by asking the learner to slow down, confirming only the first part of the description, and prompting for continuation.

Constrained production–FonFs In the constrained production system, which implemented the strongly form-focused approach, the learner's production was restricted to supplying the target form by filling a gap in a pre-scripted dialogue turn as in the example below:

(5) **S:** Wie komme ich zur Mensa? *How do I get to the cafeteria?*
L: Gehen Sie hinter ☐ Cafe nach links. *Turn left past the coffee-shop*

The learner was allowed three attempts to produce the correct form. In case an invalid form was supplied, the system signaled it with a message 'That was wrong!' and subtracted one point from a learner's "score" on the activity; correct forms increased the score by one. The feedback and the score were displayed in a designated feedback area. After the third unsuccessful attempt the correct utterance was appended to the dialogue and the system generated its next turn.

This system was built on the same architecture as the free production systems. However, due to the constraint on the language production, it used a simpler method to map the input to the expected answer; case-insensitive exact string matching. The dialogue model was also simplified because the elicitation subdialogues and more elaborate feedback mechanisms were not employed.

3 The system

All three dialogue activities have been implemented on the same system architecture. In the description in this section, we concentrate on the components required for the free production activities; the constrained production activity is its a simplified variant.

The system maintains a dialogue with the learner by following the dialogue strategy outlined in Section 2.2 (see Figure 2). This involves interpreting the learner's input, responding to the learner by selecting a communicative goal according to the dialogue model and the pedagogical strategy, and realizing the goal as a surface string. Specifically for the learning context, the system has to recognize errors in the learner input (identify contributions in the TF-realized-incorrect category) and generate feedback on them.

```
<dir-change> = Gehen Sie <pp> nach (<left> | <right>)
<pp>         = <pp-DATIVE> {dat} | <pp-NODAT> {non-dat}
<pp-DATIVE>  = <prep> <np-dat>
<pp-NODAT>   = <prep> (<np-nom> | <np-gen> | <np-acc>)
```

Figure 4: A simplified fragment of the interpretation grammar including a mal-rule; `{non-dat}` is the semantic tag indicating that a dative PP was not used where it was expected.

Figure 3 shows the system's architecture: the modules, the resources they employ and the units of information that are passed between them.

The dialogue model and engine The dialogue model represents the sequences of possible turn transitions: alternating turns produced by the user and the system. It is implemented as a state machine using State Chart XML (SCXML) as an underlying representation. We use the Java implementation of Apache SCXML.[4] The Apache framework also provides a dialogue execution engine which receives input interpretations and triggers system responses according to the model.

Interpretation of learner's input In general, interpreting the user input involves mapping a surface string of an utterance to a meaning representation. As typical in small-scale dialogue systems, we implemented the system's language model (the set of linguistic expressions it covers) as a context free grammar with semantic tags. For parsing, we use the Java Speech API implementation of the CMU parser which is part of the Sphinx system.[5] The semantic tags encode two types of information: first, the symbolic meaning of utterances, and second, information on violations of grammatical constraints. Two error handling strategies are implemented in the system:

Fuzzy matching for unknown words In order to ensure robustness with respect to typos and spelling errors the system first identifies unknown words in the input and tries to map them to known words by calculating the Levenshtein distance between the unknown word and known words. For replacement with in-vocabulary candidates we consider those words which have a Levenshtein distance within a certain range to a known word, normalized by word length.

Grammatical error handling Since the system interacts with learners, i.e. non-native speakers of German, their input is likely to contain other errors apart from misspellings, in particular errors in the target structure. An essential requirement of the system is to recognize those errors and give feedback on them. One strategy to deal with errors is to explicitly integrate anticipated errors into the grammar in the form of so called mal-rules, i.e. grammar productions which are outside of the standard rules of the given language. Erroneous utterances are parsed using mal-rules and the parse result contains information about the error. Figure 4 presents a fragment of the interpretation grammar, including mal-rules. The rule `<dir-change>` covers the utterance given in (2). If the prepositional phrase `<pp>` is not in the dative case, the semantic tag `non-dat` is returned, indicating that the dative case was required, but was not found. We encoded a set of mal-rules based on informal prior pre-testing of the system with beginner learners.

[4] `http://commons.apache.org/scxml`

[5] `http://cmusphinx.sourceforge.net`

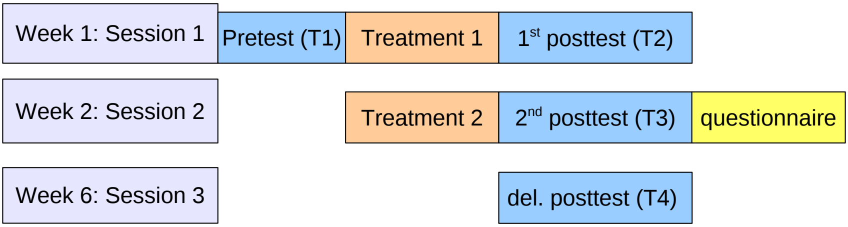

Figure 5: Experiment timeline

The drawback of this approach is that it is hard to anticipate all possible errors that might occur. Therefore, our system also implements a fall-back strategy based on keyword spotting: If no parse is found for an utterance, we create a semantic interpretation based on content words, using a keyword lexicon.

Generation of system responses The system output realization is performed using a template-based approach. The output is produced by generating a dialogue move selected according to the dialogue model using a context free generation grammar. The grammar associates atomic symbols representing communicative goals with sets of possible realizations. The generation templates contain slots encoding references to landmarks or directions for confirmation moves, or grammatical information for error feedback. Slots in the templates are filled using feature-value pairs passed as arguments to the templates along with the communicative goals to be realized.

4 The Experiment

In order to answer our research questions we conducted an in-classroom quasi-experimental study with the systems we built, in a pretest multiple-posttest design. The setup of the experiment is presented in the following sections.

4.1 Design and procedure

The study used a quasi-experimental design involving 60 students from six German language classes (ranging from A2 to B1+ CEF level (Trim et al., 2001)), taught by different teachers. The courses met twice a week for 90 minute sessions. The experiment took place six weeks (approximately 15 instruction hours) into the course. In each class, participants were randomly assigned to one of the three conditions: free production-recast, free production-metalinguistic feedback, and constrained production. Figure 5 illustrates the timeline and setup of the procedure. The experimental groups participated in two sessions of the computer-based communicative instruction with one week's break between the sessions. Each session consisted of at least two repetitions of the activity in different configurations of the task material.

At the first session, all groups completed a pretest (T1, see below) and worked with the system (the treatment) followed by a posttest (T2). Another posttest (T3) followed the second session of the treatment, and another posttest (T4) after a five week break. After the second session the subjects completed a short questionnaire on biographical information and feedback on the system.

4.2 Tests

We used two types of tests in the study: an *untimed sentence construction* test, targeting explicit knowledge, and a *timed grammaticality judgment* test, targeting implicit knowledge. Explicit knowledge is knowledge accessible through controlled processing, while implicit knowledge is accessible through automatic processing, i.e. learners' intuitive awareness of the linguistic norms (Ellis et al. (2006)).[6]

Timed grammaticality judgment Following Ellis (2006) we designed a timed grammaticality judgment test to measure implicit knowledge. The test items included different combinations of six different spatial prepositions (*bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *vor*, ('in front of'), *zu* ('to'), *auf* ('on')) and nouns of the three genders, equally balanced. The underlying problem with testing the dative case is that learners need to know the gender of the noun in order to make a judgment about the correctness of a prepositional phrase. Because we did not want to test the learners' knowledge of genders, but their knowledge of the datives, we chose common feminine and masculine nouns whose grammatical gender matches the semantic gender, e.g. mother, man, son, etc. For neuter nouns we chose words that are usually taught at the beginner's level, e.g. child, horse. However, due to logistic constraints, we could not explicitly test whether the gender of the nouns included in the test items were indeed known.

The test included 9 grammatical, 8 ungrammatical test items[7] and 7 grammatical and 7 ungrammatical distractor items. The time-limit was set to 10 seconds per item. (This is roughly twice the maximum time a native speaker used).[8] Each correctly judged item was scored at 1 point, each incorrectly judged item was scored at 0.

Sentence construction For the explicit knowledge test, participants were asked to complete sentences given the beginning of a sentence and a set of unordered uninflected phrases or words. Full noun phrases were given along with gender information, as in the example below:

Item:	Das Pferd (stehen, die Kuh, vor)
Solution:	Das Pferd steht vor der Kuh. *The horse stands in front of the cow.*

The test consisted of 8 test items containing 6 prepositions (*bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *vor*, ('in front of'), *zu* ('to'), *zwischen*, ('between')) with a gender-balanced set of nouns, and 4 distractor items.[9] There was no time-limit on the test items. The item was scored 1

[6] The tests were prepared and administered using Webexp Experimental Software. http://www.hcrc.ed.ac.uk/web_exp/

[7] One of the original 9 ungrammatical items was disregarded in the evaluation because of a spelling error we overlooked.

[8] Ellis timed his test at 20% above the average time native speakers required (Ellis, 2006). Han and Ellis (1998) used 3.5 seconds as the time constraint based on pretesting the items, while Bialystok (1979) used an even shorter time limit. Based on our pretest, already the threshold of 3.5 seconds would have excluded a couple of slow native speakers. Since we are not aware of research which explicitly addresses the issue of the time limit on the timed judgment tasks, we opted for a more generous time-limit.

[9] Note that the used prepositions differ slightly between the two tests types for practical reasons: For instance, although 'between' is a relevant preposition, we did not use it in the grammaticality judgment test, because it requires two noun

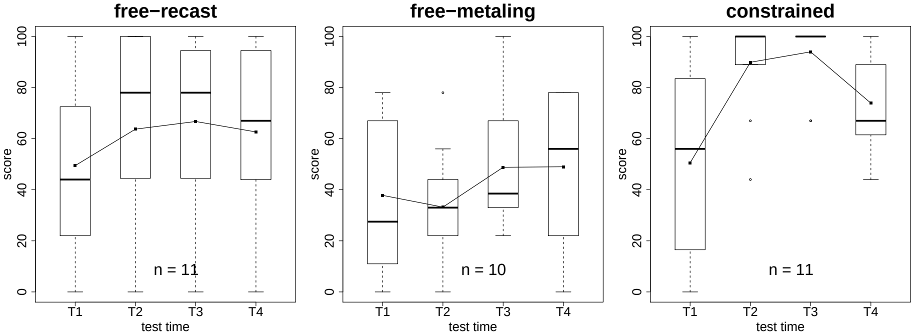

Figure 6: Results for sentence construction test.

point if the prepositional phrase was built correctly. The item with the preposition ‘between’ was scored at 1 point for each correct noun phrase. All other form errors were neglected.

We created four versions of each of the tests to be administered at the four times of assessment (T1, T2, T3, T4). The versions differed in the combinations of prepositions and noun phrases, but were otherwise comparable with regard to lexical items used. The assignment of a test version to a test time was randomly varied for each participant in order to compensate for any unintended differences between test versions. Within each test items were presented in random order.

4.3 Analysis

With the experiment spanning over six weeks, subject drop-out was inevitable. Due to a high drop-out rate (around 50%) and incidental data loss we have data for only 32 subjects for the sentence construction test, and 30 subjects for the grammaticality judgment test, around 10 for each experiment condition. We performed non-parametric analyses because of the small sample size and because parametric assumptions were not met: According to Shapiro-Wilk and Levene tests, both the normality assumption and the assumption of homogeneity of variance were violated on at least some of the within-subject and/or between-subject variables on either tests.

In order to compare within subject differences we performed Friedman tests followed by pairwise post-hoc comparisons using Wilcoxon signed rank test on those groups for which the Friedman test was statistically significant. For between-group comparisons we used the Kruskal-Wallis one-way analysis of variance test, followed by pairwise post-hoc comparisons using the Mann-Whitney U test. The significance level was set at 0.05. We mark differences which were significant at $\alpha = 0.10$ to indicate interesting tendencies.

		T1		T2		T3		T4	
Group/test	n	m	sd	m	sd	m	sd	m	sd
Free-recast									
SC total	11	49	32	64	39	67	35	63	38
TGJT total	11	63	23	78	22	79	18	76	17
TGJT gram.		76	21	83	18	84	15	88	13
TGJT ungr.		49	29	73	33	73	27	62	31
Free-metaling									
SC total	10	38	31	33	22	49	26	49	32
TGJT total	9	52	18	62	22	63	20	64	15
TGJT gram.		63	22	77	22	78	16	83	14
TGJT ungr.		40	29	46	33	46	34	43	27
Constrained									
SC total	11	51	36	90	18	94	13	74	19
TGJT total	10	68	19	90	16	87	13	81	18
TGJT gram.		78	21	92	10	89	12	89	14
TGJT ungr.		58	27	88	24	85	18	71	26

Table 2: Test results: means (m) and standard deviations (sd) for percentage scores

5 Results

The analyses below are based on the data set for the 30 (or 32) subjects with test results for all four assessment times. Table 2 shows the mean percentage scores and standard deviations for each experimental group on both tests: sentence construction (SC) and timed grammaticality judgment test (TGJT). For the latter test, the table also shows the scores for grammatical and ungrammatical items separately.

5.1 Sentence construction test

Figure 6 shows box plots and means for the sentence construction test. The first point to note is that there was no significant between group difference on the pretest on both tests. This means that before the treatment the groups were at the same level. The free-recast group and the constrained production group both increased from pretest (T1) to the first posttest (T2) and slightly further increased between from the first posttest (T2) to the second posttest (T3). The free-metalinguistic group slightly deteriorated between T1 and T2, but improved between T2 and T3. All groups deteriorated at the delayed posttest (T4). Within-subject analysis of variance showed that there were significant differences in the scores across the three time periods in the constrained production

phrases that have to be judged at the same time, which makes it impossible to determine based on which the judgment was made.

group, but not in the other conditions. Post-hoc analysis showed that the constrained production group was significantly more accurate on the two posttests (T2 and T3) than on the pretest (T1).

Between-group comparisons showed significant difference at T2 and T3. Post-hoc analyses showed that at T2, the free-recast and the constrained production group were both significantly more accurate than the free-metalinguistic group.

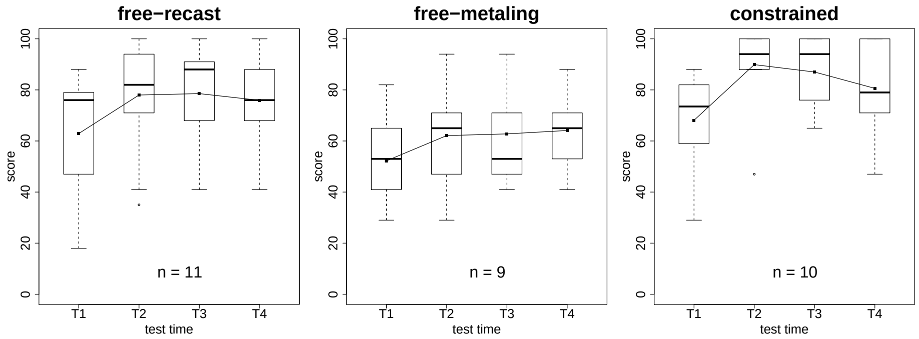

Figure 7: Results for timed grammaticality judgment test.

5.2 Timed grammaticality judgment test

Figure 7 illustrates the results for the timed grammaticality judgment test. As with the sentence construction test, the groups showed no significant difference at T1. All groups increased between T1 and T2. Between T2 and T3, only the constrained production group slightly deteriorated. The other groups showed no difference. All groups slightly deteriorated between T3 and T4. Within-subject analysis of variance showed that there were significant differences in the scores across the three time periods in the constrained production condition, with post-hoc analysis showing that the accuracy was significantly higher on T2, T3 and T4 than on T1. The free-recast group showed a difference across all test times that was significant at $\alpha = 0.10$, and further analysis showed that this group yielded a significantly higher score at T2 and T3 than at T1, ($p < 0.05$). The free-metalinguistic group showed no significant difference across test times. If we consider only the ungrammatical items of the grammaticality judgment test, these differences are maintained. However, for the grammatical items, the free-metalinguistic group shows a marginally significant difference ($p = 0.06$), with post-hoc analysis revealing that their score at T3 is higher than on T1 (marginally significant, $p = 0.06$) and their score at T4 is higher than on T1 ($p < 0.05$). The other two groups show no significant within-subject differences on the grammatical items.

Between group comparisons showed that there are marginally significant differences between groups at T2 and T3. Post-hoc analysis revealed that in both instances, the free-recast and constrained production group scored higher than the free-metalinguistic group.

groups	free-recast			free-metaling			constrained		
tests	●●○○	●●●○	●●●●	●●○○	●●●○	●●●●	●●○○	●●●○	●●●●
n: SC/TGJT	21/21	18/18	11/11	25/25	20/20	10/9	20/21	14/15	11/10
SC-total	(1-2)	-	-	-	1-3,(2-3)	-	1-2	1-2,1-3	1-2,1-3
TGJT-total	1-2	-	1-2,1-3,1-4	1-2	1-2,(1-3)	-	1-2	1-2,1-3	1-2,1-3
TGJT-gram.	-	-	-	-	-	(1-3),1-4	1-2	1-2	-
TGJT-ungr.	1-2	1-2	1-2,1-3	1-2	1-2	-	1-2	1-2,1-3	1-2,1-3,(1-4)

Table 3: Significant changes between test times for each group and each subset of tests taken.

5.3 Additional results for subsets of assessment times

Given the small number of subjects leading to low power of the tests, we also conducted analyses with data comprising only the first two or three tests respectively, for which the data of 67 (or 53, respectively) subjects is available. Table 3 shows for which subsets of assessment times there were significant within- and between-subject differences. For each of the three conditions the table shows three columns indicating the data of all subjects taking part in the first two (●●○○), the first three (●●●○) , or all four tests (●●●●) respectively. The values in the table cells indicate between which of the respective tests there was a significant difference. Brackets indicate that the difference is only significant at $\alpha = 0.10$. For instance, for the free-metalinguistic group on the timed grammaticality judgment test (TGJT-total), for all subjects taking part in all tests (●●●●) , there was a significant difference between T1 and T4 (at $\alpha = 0.05$) and a marginally significant difference between T1 and T3 (at $\alpha = 0.10$) .

While we found similar within-subject differences for the constrained production group, most interestingly, more differences became evident for the free production groups. More specifically the free-metalinguistic group showed significant improvement on the sentence construction test between T1 and T3 (and between T2 and T3 at $\alpha = 0.10$). This group also showed an increase in accuracy on ungrammatical items between T1 and T2. The free-recast group showed an increase in performance in the sentence construction test between T1 and T2 (at $\alpha = 0.10$) when only considering data for these two assessment times.

6 Conclusion

We presented a study which investigated the efficacy of different computer-based form-focused task-oriented activities on the acquisition of the German dative in a certain type of prepositional phrases. Noting that the number of subjects whose data we were able to analyze statistically was rather small, the implications of this study should be taken cautiously. Based on the analyses, certain tendencies can be however observed.

First, not surprisingly, most of of the effect is found between the pretest and the first posttest, that is, there is an immediate effect of the intervention. Second, also not surprisingly, the explicit Focus on Forms instruction (constrained production) appears to achieve more of the effect.[10] It

[10] Considering the drill-like character of the constrained production dialogues, it would be of course interesting to contrast it with a simple traditional decontextualized drill in order to see whether there is any added value to the embedding in the dialogue interaction.

appears that the learning in the free production conditions is slower (stepwise increase in the mean scores in the free production groups vs. a jump of the scores in the constrained condition). We cannot draw a clear conclusion to our third research question, whether there is a difference between the different feedback types. In general, the recast group achieves more significant gains in accuracy when taking into account the data for all four test times. However, if we do not consider the delayed posttest, the metalinguistic group seems to achieve the same effect on a larger data set.

It is interesting that the free-recast group achieves more of the significant results on the implicit knowledge test than on explicit knowledge. This might be due to, on the one hand, indirect nature of the feedback and a weaker form-focusing mechanism than in the other condition, and on the other hand, due to stronger engagement in the activity and, possibly, better noticing of feedback (recasts) as a result.

While the presented analysis focused on accuracy in the usage of the target structure as the only measure of language development, we also tested the effect of the task activities on spoken language fluency on an analogous task. In the beginning of each session participants were asked to work in pairs and describe a route on a map. The ensuing conversations were recorded. For the subset of the data – 13 participants of the constrained production and the free-recast condition – we analyzed the transcripts of the speech samples with regard to durational measures associated with fluency. In addition, we also asked German teachers to rate and rank those samples with respect to the perceived fluency. However, the results were not clear cut. When correlating the ranking of raters with the test times, a slightly higher positive correlation was evident for the free-recast group than for the constrained production group. On some durational measures, the free-recast groups improved significantly while the constrained production group showed no difference. For other measures it was the other way round. However, given the small number of subjects, again these results have to be taken cautiously.

As part of future work, we are planning to analyze the accuracy in the use of the target structure within the oral test as well as in the system interaction dialogues. We are presently annotating the interaction data (the system logs) along two dimensions: the grammatical aspects of the learner language and the structure of the interaction, in order to be able to investigate interaction-based correlates of the results we presented in this paper.

Acknowledgments

We would like to thank Dr. Kristin Stezano Cotelo and Meike van Hoorn, language instructors at the German courses of the International Office at the Saarland University, as well as their students who participated in the experiments, for the help in conducting this study.

Sabrina Wilske's work was funded by the IRTG PhD program "Language Technology and Cognitive Systems". Magdalena Wolska's position at Saarland University is partially funded through the INTERREG IV A programme[11] project ALLEGRO (Project No.: 67 SMLW 1 1 137).

[11] http://www.interreg-4agr.eu/

References

Bialystok, E. (1979). Explicit and implicit judgements of l2 grammaticality. *Language Learning*, 29:81 – 103.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, 27(3):431–463.

Ellis, R., Loewen, S., and Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of l2 grammar. *Studies in Second Language Acquisition*, 28:339–368.

Ferreira, A. (2006). An experimental study of effective feedback strategies for intelligent tutorial systems for foreign language. In *IBERAMIA-SBIA*, pages 27–36.

Han, Y. and Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, 2:1–23.

Harless, W., Zier, M., and Duncan, R. (1999). Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *Calico Journal*, 16(3):313–37.

Holland, V. M., Kaplan, J. D., and Sabol, M. A. (1998). Preliminary tests of language learning in a speech-interactive graphics microworld. *Calico Journal*, 16(3):339–359.

Johnson, W., Marsella, S., and Vilhjaálmsson, H. (2004). The DARWARS tactical language training system. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Krashen, S. D. (1985). *Input Hypothesis: Issues and Implications*. Longman.

Long, M. H. (1981). Input, interaction and second language acquisition. In Winitz, H., editor, *Native language and foreign language acquisition*, volume 379, pages 259–78. Annals of the New York Academy of Sciences.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In Bot, K., Ginsberg, R. B., and Kramsch, C., editors, *Foreign language research in cross-cultural perspective*, pages 39–52. John Benjamins.

Lyster, R. and Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19:37–66.

Norris, J. M. and Ortega, L. (2001). Does Type of Instruction Make a Difference? Substantive Findings From a Meta-analytic Review. In Ellis, R., editor, *Form-Focused Instruction and Second Language Learning*, pages 157–213. Blackwell.

Petersen, K. (2010). *Implicit corrective feedback in computer-guided interaction: Does Mode Matter?* PhD thesis, Georgetown University.

Seneff, S., Wang, C., and Zhang, J. (2004). Spoken Conversational Interaction for Language Learning. In *Proceedings of INSTIL/CALL*, pages 151–154, Venice, Italy.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Gass, S. and Madden, C., editors, *Input in second language acquisition*, pages 235–53. Newbury House.

Trim, J., North, B., and Coste, D. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Niveau A1, A2, B1, B2*. Langenscheidt.