Form-focused task-oriented dialogues for computer assisted language learning: A pilot study on German dative

Magdalena Wolska, Sabrina Wilske

Computational Linguistics, Saarland University, Saarbrücken, Germany
{magda,sw}@coli.uni-saarland.de

Abstract

We report on a pilot experiment conducted in order to investigate whether computer-based conversational focused tasks promote acquisition of forms. The structure we targeted was the German dative case in prepositional phrases. The goal of the task we designed was two-fold: First, learners should improve their overall communicative skills in the scenario and, second, expand their mastery of the target structure. In this paper, we present an evaluation of learners' progress on the latter.

Index Terms: computer-assisted language learning, computer-based form-focused dialogue tasks, German

1. Motivation

Conversational interaction is for the foreign language learner the ultimate site of language acquisition. It involves negotation of meaning, which may arise from communication problems, it is a source of comprehensible, useful input on the one hand, and, on the other hand, of negative feedback delivered by means of reformulations or clarification questions. Crucially, conversational interaction is an opportunity for learners to produce comprehensible output as well as to modify their output in response to feedback, thereby stimulating learning [1, 2, 3].

The communicative approch, based to a large extent on participatory interaction, advocates the use of goal-oriented communicative activities, *tasks*, in foreign language teaching [4]. Communicative goals of tasks should be framed in real world situations and encourage learners to use their developing language. Important definitional properties of tasks are (1) primary focus on meaning, (2) clearly defined communicative outcome, and (3) free use of linguistic forms. The third point gives rise to a potential problem when, as part of 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 form-focused instruction and the communicative approach. Focused tasks are designed in such way that learners are likely to use a specific target structure thereby improving its mastery.

We report on a preliminary experiment conducted in order to find out whether *computer-based* focused dialogue tasks also promote acquisition of forms. The structure we targeted was the German dative case in locative prepositional phrases. The goal of the task was two-fold: learners should improve their overall communicative skills in the scenario and their mastery of the target structure. In this paper, we report results on the latter.

The idea of computer-based dialogue activites for foreign langauge learning is not new. With the progress in language technology, the number of intelligent computer-assisted language learning (CALL) systems which allow learners to use natural dialogue has been growing. A CALL system can be evaluated in terms of learning gains it generates [5, 6], usability (How did learners enjoy playing with the system? [7, 8, 9]) or its performance from an engineering perspective (Did it fail? [10]). We built a CALL system which implements an established SLA methodology (focused tasks) and evaluated whether iteracting with the system we built produces learning gain.

Outline In Section 2 we present the setup of the study. In Section 3 we present the design of the pilot experiment we conducted as well as the language tests we used in the evaluation. In Section 4 we present the results of the study and conclude with a discussion and directions for further work in Section 5.

2. The approach

We designed a study to investigate the effect of computer-based form-focused task-oriented communicative activites on learning foreign language structures. Our research questions were the following: (i) Do computer-based form-focused dialogue tasks help learners of German improve accuracy on dative in locative prepositional phrases? (ii) Is there a difference in the effect of free vs. constrained type-written production on learning.

2.1. The target form and the task

For the focused communicative activity we selected a grammatical form and a task with the following properties: (1) We wanted a structure with a scenario which creates incidental opportunites for the learner to produce the stucture because it is natural to use, (2) it has enough distinguishing features to be easily tested, (3) the scenario and the communicative task are meaningful, realistic, and useful for the learner. With this in mind, we chose an activity focused on the German dative in prepositional phrases (PP) and framed in the context of a "Directions giving" dialogue. We introduce the parameters below.

Form: Dative case in locative prepositional phrases The dative case in German is required as an object of certain spatial prepositions, certain verbs govern the dative and it typically marks the indirect object.[1] The case 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 too).

Prepositions requiring dative include, among others, *nach* ('at', 'past'), *hinter* ('behind'), or *zwischen* 'between'. Most locative prepositions used for describing static spatial relations require dative, as does the directional preposition *(bis) zu* ('to'). All these can be elicited in a task which requires spatial descriptions and directions. An obvious scenario, then, in which a task-based activity can be embedded is the "Directions giving" scenario, often employed in course-books for this purpose.[2]

[1] We do not address these latter two uses in this experiment.

[2] Note that, according to native speaker judgement, while Accusative

Task: Directions giving We designed a directions giving task in such way that it 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.[3] The map has two turning points and a landmark at each as well as at the target point; these create opportunities to use PPs to describe the point of turn. We balanced the landmarks as to gender and provided the genders on the map thus eliminating a confound in case a subject should not know them.[4] The route includes two points of direction change with landmarks around them.[5]

2.2. Task-based activities

We designed and implemented 2 variants of a role play activity framed within the task described above: In both variants it involved a (type-written) dialogue with the system we built in order to perform "Directions giving" task. The system controlled the interaction by means of a state-based dialogue model and explicitly implemented form-focusing mechanisms (as part of dialogue model or by restricting the input mode). The dialogue model encoded subdialogues for *eliciting target forms* and *provided feedback on forms* in case the learner supplied them incorrectly. The activities differed in the freedom of language production and the realisation of form-focused feedback:

Free language production In free-production, learners were able to type utterances freely without restrictions on language. The system implemented two input interpretation strategies: one based on a grammar with mal-rules,[6] and a fall-back based on fuzzy keyword matching. The system classified the learner's input into one of three categories ("TF": "target form"): TF-realized-correct, TF-realized-incorrect, TF-not-realized.

The high-level dialogue and feedback strategy of the system can be summarised as follows: If the learner's utterance could be parsed, the system provided implicit feedback in case of learner errors in the TF by reformulating (*recasting*) the learner's utterance (or parts thereof). Recasts were realised in a way so as to give them an appearance of implicit confirmation type of grounding moves (see (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 (2). The two dialogue situations are exemplified below:[7]

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

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

The system did not attempt to diagnose nor correct any other incorrect structures except those in focus.[8]

Constrained production In the constrained system the learner's production was restricted to supplying the target form by filling a gap in a pre-scripted dialogue turn as shown below:

(3) **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'

Subjects were allowed 3 attempts to produce the correct form. In case an invalid form was supplied, the system signalled it with a message 'That was wrong!' and subtracted 1 point from a learner's "score"; correct forms increased the score by 1. The feedback and the score were displayed in a designated area. After the third unsuccessful attempt correct utterance was appended to the dialogue. The system then generated its next turn based on the dialogue model. The system was built on the same architecture as the free production system, however, due to the constraint on 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 elicitation subdialogues and recasts were not employed.

3. Experiment

We conducted an in-classroom experiment with the systems we built in a nonrandomized pretest multiple-posttest design. The setup of the experiment is presented in the following sections.

Design, participants, procedure The study used a quasi-experimental design involving 22 students from two German language classes at the univerity, taught by different teachers. The classes were split randomly into two sub-groups: one assigned to the free production condition, and the other to the constrained production condition. The participants came from different language backgrounds, were both male and female, with an average age of 24 years, and had been learning German for an average of 2 years prior to experiment.[9] The courses met twice a week for 90 minute sessions. The experiment took place 6 weeks (ca. 15 instruction hours) into the course.

The two experimental groups participated in two sessions of form-focused task-based activites 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 (the map) as described in Section 2.1. To complete the activity the participants took between 5 and 24 minutes (13 mins on average) in the free condition and between ca. 2 and 10 minutes (average 4:30 mins) in the constrained condition.

At the first session, the groups completed a pretest and an immediate post-test (posttest1). The groups completed another post-test at the end of the second session (posttest2), and a another post-test after five week's break (delayed posttest); see below. After the second session the participants filled out a short

is also possible in certain directions giving contexts, its use in the scenario we exploit would be marked and does not typically occur.

[3] The instructions state that they were stopped on the campus and asked for directions. We explicitly request that the map provided be used and that the indicated route be described. The task description does not, however, contain any explicit hint to use PPs or the dative.

[4] We did not include streets or crossroads in order to prevent reference to those. Initial testing with more realistic maps had shown that learners prefer to refer to crossings and streets, thus effectively avoiding the target structure. While these can also be used in dative PPs, 'crossing' and 'street' are both feminine, so the opportunities to build dative PPs with masculine and neuter nouns would be reduced.

[5] Participants repeated the activity twice. For the second round, the map was kept, but start point, goal, route and landmarks were changed.

[6] We use a simple grammar analogous to those employed in spoken dialogue systems. The grammar was encoded in the Java Speech Grammar Format and parsed using an open-source parser extracted from the CMU Sphinx system; `http://cmusphinx.sourceforge.net`.

[7] **S** and **L** mark system and learner turns. Bold did not appear in the output; here it indicates the incorrect form and correction via recast.

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

[9] Their German proficiency level was classified as ranging from A2 to B1+ CEF level, based on scores on an initial course placement test.

questionnaire giving biographical information and feedback on the interaction with the system.

Tests We used two types of tests: an **untimed sentence construction test**, targeting explicit knowledge and a **timed grammaticality judgment test**, targeting implicit knowledge.[10]

To test explicit knowledge, the participants were asked to build sentences given the beginning of a sentence and a set of unordered uninflected phrases or words. (Full noun phrases with the gender information were privided.) The test consisted of 8 items with 6 PPs[11] and 4 distractors. There was no time-limit. The items were scored at 1 if the PP was built correctly. The item with preposition 'between' was scored at 1 point for each correct noun phrase. All other form errors were neglected. The maxium score for the sentence construction test was 8.

The timed grammaticality judgment test was designed following Ellis [12] and included 9 grammatical, 8 ungrammatical items and 7 grammatical and 7 ungrammatical distractors.[12] The time-limit was set at 10 sec. per item. This is roughly twice the maxium time a native speaker used.[13] Each correctly judged item was scored at 1, 0 otherwise; maximum score was 9.

We created four versions of each of the tests for the four times of assessment (pretest, posttest1, postest2, and delayed postest). 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 varied for each participant in order to compensate for any unintended differences between test versions. Within each test, items were presented in random order. Both groups used the same tests and were scored in the same way.

Analysis With the experiment spanning over six weeks, subject drop-out was inevitable. Due to a high drop-out rate (over 50%) we have data for only 12 subjects for all the four assessment points for both tests. Because of the small sample size and because parametric assumptions were not met,[14] we performed non-parametric analyses. 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.

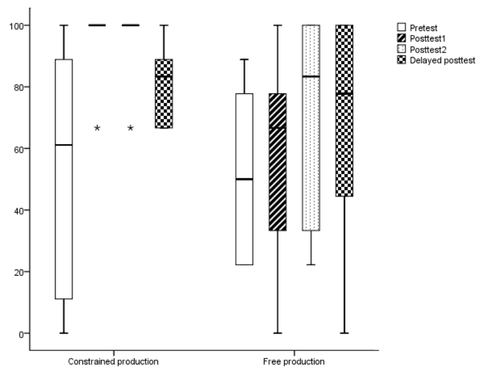

Figure 1: Sentence construction results

		Pretest		Posttest 1		Posttest 2		Delayed Posttest	
Groups	N	M	SD	M	SD	M	SD	M	SD
Constrained	6	53.70	40.62	94.44	13.61	94.44	13.61	81.48	13.46
Free	6	51.85	29.54	57.41	36.12	70.37	34.19	66.67	39.13

Table 1: Numerical results of the sentence construction test: means (M) and standard deviations (SD) for percentage scores

For between-group comparisons we used the Mann-Whitney U test. The significance level was set at 0.05.

4. Results

All the analyses below are based on the data set for the 12 subjects with test results for all the four assessment times: 6 in each of the two treatment groups.

Sentence construction test Table 1 shows the percentage scores' means and standard deviations (see also Figure 1). The first point to note is that there was no significant between group difference on the pretest, according to the Mann-Whitney U test. This means that before the treatment the groups were at the same level. Both groups increased accuracy of dative use from pretest to posttest1, however, while the free production group further increased between posttest1 and posttest2, the constrained group stayed at the same level. In both groups accuracy declined between posttest2 and delayed posttest.

Within-subjects analysis of variance showed that there were significant differences in the scores across the three time periods in the constrained production condition ($\chi^2 = 11.69$, df$= 3$, $p < 0.05$). Post hoc analysis showed that the constrained production group was significantly more accurate on posttest1 than on pretest ($Z = -2.02$, $p < 0.05$) and marginally declined from posttest2 to delayed posttest ($Z = -1.84$, $p = 0.07$). In the free production group, however, the difference across the times of assessment was not significant according to the level we had set ($\chi^2 = 5.43$, df$= 3$, $p = 0.14$). Between-group comparisons using the Mann-Whitney U test show significant difference on posttest1 ($U = 5.50$, $Z = -2.14$, $p < 0.05$), but no significant differences on the other two posttests.

Grammaticality judgement test Table 2 shows the percentage scores' means and standard deviations (see also Figure 2). There was no significant between group difference on the pretest; neither on total scores nor on grammatical and ungrammatical item scores. We noted, however, a few extreme outliers. Therefore, we interpret the following results with caution.

In general, based on descriptive statistics both groups increased accuracy from pretest to posttest1. There was a significant within-subjects difference in the total scores across the

[10] Explicit knowledge is knowledge accessible through controlled processing, while implicit knowledge refers to knowledge accessible through automatic processing and which learners are intuitively aware of [11]. The tests were prepared and administered using Webexp Experimental Software. `http://www.hcrc.ed.ac.uk/web_exp/`

[11] *zu* ('to'), *vor*, ('in front of'), *bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *zwischen*, 'between'; The set of nouns was gender-balanced.

[12] The targets were combinations of 6 spatial prepositions and nouns of the three genders, equally balanced. The prepositions were: *auf* ('on'), *bei* ('at'), *hinter* ('behind'), *neben* ('next to'), *vor*, ('in front of'), *zu* ('to'). The problem with testing dative is that learners need to know the gender of the noun. Because we did not want to test the learners' knowledge of gender, 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 usually taught at the beginnger's level, e.g. child, horse. However, we did not explicitly test whether the genders of the nouns were indeed known.

[13] Ellis timed his test at 20% above the average time native speakers required [12]. Han and Ellis used 3.5 seconds as the time constraint in [13] based on pretesting the items, while Bialystok used an even shorter time limit [14]. Based on pretesting, already a 3.5 second threshold 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 judgement tasks, we opted for a more generous time-limit.

[14] According to Shapiro-Wilk and Levene tests, both the normality and the homogeneity of variance assumptions were violated on at least some within-subject and/or between-subject variables on either tests.

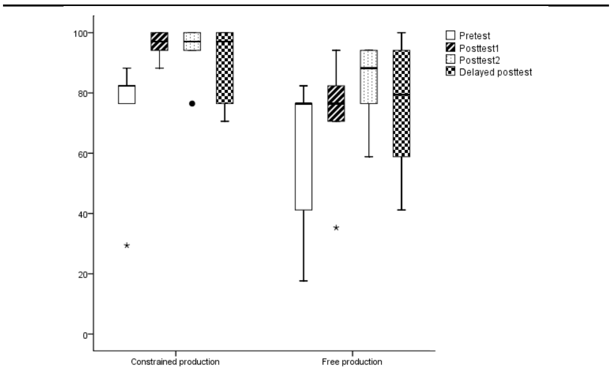

Figure 2: Grammaticality judgement results: totals

Groups	N	Pretest		Posttest 1		Posttest 2		Delayed Posttest	
		M	SD	M	SD	M	SD	M	SD
Constrained	6								
Total		73.53	21.93	96.08	4.80	94.12	9.11	90.20	13.24
Grammatical		75.92	26.68	96.30	5.74	96.30	5.74	94.44	9.30
Ungrammatical		70.83	18.82	95.83	10.20	91.67	15.14	85.42	20.03
Free	6								
Total		61.76	26.24	72.55	20.26	83.33	13.63	75.49	22.14
Grammatical		74.07	25.01	75.92	20.39	85.18	11.47	87.04	17.80
Ungrammatical		47.92	30.02	68.75	24.68	81.25	18.96	62.50	36.23

Table 2: Numerical results of the grammaticality judgment test

three time points in both conditions (constrained condition: $\chi^2 = 11.18$, df$=3$, $p < 0.05$; free: $\chi^2 = 8.12$, df$=3$, $p < 0.05$). Post hoc analyses showed an increase between pretest and postest1 ($Z = -2.26$, $p < 0.05$). Between-group comparisons showed statistically significant difference on posttest1 ($U = 2.00$, $Z = -2.61$, $p < 0.05$) and marginally significant difference on posttest2 ($U = 6.50$, $Z = -1.89$, $p = 0.06$), but no significant differences on the pretest nor delayed posttest. Interestingly, analysis of the grammatically correct and incorrect items individually showed a marginal difference across the time periods on both grammatical and ungrammatical items in the constrained condition (grammatical items: $\chi^2 = 7.50$, df$=3$, $p = 0.06$; ungrammatical: $\chi^2 = 11.94$, df$=3$, $p < 0.05$) as well as on the ungrammatical items in the free production condition ($\chi^2 = 8.67$, df$=3$, $p < 0.05$). At this point, however, post hoc analysis may be misleading because of extreme outliers present in the data.

5. Discussion and future work

We consider the completed study as only preliminary: firstly, because the number of subjects whose data we were able to analyse statistically was very small and secondly, because the distribution of the data for the judgement test was heavily skewed by outliers, making it difficult to draw firm conclusions.

Based on the analyses, certian tendencies can be however observed. First, not surprisingly, most of of the effect is found between the pretest and the posttest1, i.e. there is an immediate effect of the intervention. Second, also not surprisingly, the strongly form-focused interaction (constrained) appears to achieve more of the effect. However, some of the comparisons we originally made at the significance level of 0.05 were in fact significant at a more liberal level of 0.10; also those in free production condition. Third, it appears that learning in the free production is slower (stepwise increase in the mean scores vs. a jump of the scores in the constrained condition; see boxplots). It would be interesting to see whether the benefit of this condition can be found in other language skills which might be supported indirectly. In fact, aside from the two assessment tests we also asked the learners to take two oral tests which consisted of another variant of the same role-play task, only performed in spoken dialogues. What we would like to investigate here is whether the forms and the general conversational skills acquired in the given scenario transfer to oral production and if so whether there is a difference in the effect between the two conditions. Finally, it is interesting that the free production group achieves more of the significant results on the implicit knowledge. This might be due to, on the one hand, the indirect nature of freedback and a weaker form-focus mechanism than in the other condition, and on the other hand, due to stronger engagement in the activity and the resulting better noticing of recasts. We are presently coding the dialogues in order to investigate interaction-based correlates of the results.

6. Acknowledgements

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. Magdalena Wolska's position at Saarland University is funded through the INTERREG IV A project "ALLEGRO" (project number: 67 SMLW 1 1 137).

7. References

[1] M. H. Long, "Input, interaction and second language acquisition," in *Native language and foreign language acquisition*. Annals of the New York Academy of Sciences, 1981, vol. 379, pp. 259–78.

[2] S. D. Krashen, *Input Hypothesis: Issues and Implications*. London: Longman, 1985.

[3] M. Swain, "Three functions of output in second language learning," in *Principle and practice in applied linguistics*. Oxford: Oxford University Press, 1995, pp. 125–144.

[4] R. Ellis, *Task-based Language Learning and Teaching*. Oxford University Press, 2003.

[5] V. M. Holland, J. D. Kaplan, and M. A. Sabol, "Preliminary tests of language learning in a speech-interactive graphics microworld," *Calico Journal*, vol. 16, no. 3, pp. 339–359, 1998.

[6] W. Harless, M. Zier, and R. Duncan, "Virtual dialogues with native speakers: The evaluation of an interactive multimedia method," *Calico Journal*, vol. 16, no. 3, pp. 313–37, 1999.

[7] C. Wang and S. Seneff, "A spoken translation game for second language learning," in *Proc. of AIED-07*, 2007, pp. 315–322.

[8] T. Lech and K. de Smedt, "Dreistadt: A language enabled MOO for language learning," in *Proc. of the ECAI-06 Workshop on Language-enabled Educational Technology*, 2006, pp. 38–44.

[9] W. L. Johnson and S. Wu, "Assessing aptitude for learning with a serious game for foreign language and culture," in *Proc. of Intelligent Tutoring Systems Conference*, 2008, pp. 520–529.

[10] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *Proc. of InSTIL*, 2004, pp. 151–154.

[11] R. Ellis, S. Loewen, and R. Erlam, "Implicit and explicit corrective feedback and the acquisition of l2 grammar," *Studies in Second Language Acquisition*, vol. 28, pp. 339–368, 2006.

[12] R. Ellis, "Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge," *Applied Linguistics*, vol. 27, no. 3, p. 43163, 2006.

[13] Y. Han and R. Ellis, "Implicit knowledge, explicit knowledge and general language proficiency," *Language Teaching Research*, vol. 2, pp. 1–23, 1998.

[14] E. Bialystok, "Explicit and implicit judgements of l2 grammaticality," *Language Learning*, vol. 29, pp. 81 – 103, 1979.