ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Assessing the Effect of Type-written Form-focused Dialogues on Spoken Language Fluency

Magdalena Wolska, Sabrina Wilske

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

Abstract

We report on a preliminary study which investigated the effect of form vs. meaning focused computer-based *type-written* activities on *spoken* communicative skills of learners of German. Both types of activities were realized as written interactions with a computer-based dialogue system and were framed within a "Directions giving" scenario. The linguistic form in focus was the German dative case in prepositional phrases. The paper introduces the methodology we adopt to assess the learners' language fluency (holistic fluency ratings and temporal measures of speech) and reports on the results of a pilot experiment.
Index Terms: computer-assisted task-based language learning, second-language acquisition, communicative approach, fluency

1. MOTIVATION

Foreign language instruction can give priority to the formal aspects of language or to meaning and content. The distinction between emphasis on form versus meaning is often referred to as Focus on Forms (FonFS) vs. Focus on Form (FonF), the core difference being that FonFS addresses linguistic forms in isolation, providing no or limited meaningful context, while FonF tries to integrate meaning and form by drawing learners' attention to forms as they arise within primary meaning-based language learning activities. The degree to which instruction emphasises linguistic forms can vary: explicit instruction explains the language phenomenon in question or asks learners to attend to particular forms in the target language. The key aspect of the FonF methodology is to provide learners with incidental opportunities to produce comprehensible output as well as to modify their output in response to content-motivated (implicit) form-focused feedback [1, 2]. This can be realized within focused communicative activites such as *focused tasks*, that is, tasks designed in such way that learners are likely to use a specific target structure, without being told what the specific linguistic focus is and without explicit emphasis on forms [3].

To date there appears to be no consensus as to what type of instruction is most effective in foreign language teaching, in particular, implementing the communicative approach. Several comparative classroom studies of explicit vs. implicit instruction and FonFS vs. FonF have been conducted resulting in mixed conclusions; see [4] for an overview on form-focused instruction and [5] for an overview on meaning-oriented focus on form. An obvious question to ask in the context of computer-assisted language learning (CALL) is whether *computer-based* task activities have an effect on language acquisition. While spoken-input CALL systems do exist (see, for instance, [6, 7]) spoken learner language is in general hard to analyze automatically, especially when subtle phonetic differences are involved in form-focused free production; however, see [8, 9] for recent advances in non-native speech recognition. It is therefore all the more interesting to investigate the effect of *type-written* computer-based activites on *spoken* language production.

The paper reports on a preliminary study which investigated the effects of explicit (FonFS) vs. implicit (FonF) type-written communicative focused activities on spoken fluency. The activities were set in the "Directions giving" scenario. The FonF activity involved free production dialogue during which a German CALL system we built provided *incidental focus on form* by *implicitly correcting* learner errors in the dative case. While our research addressed two issues: the effect of computer-based FonFS instruction vs. FonF on (i) the accuracy in the use of forms, and (ii) the spoken production in a communicative task, in this paper we present results on the latter question: *Do computer-based form-focused type-written dialogues have an effect on spoken production fluency?*[1] To our knowledge no analogous experiments have been previously conducted.

2. FORM-FOCUSED DIALOGUES

In order to investigate the research question formulated above, we set up an experiment in which language learners interacted with one of two variants of a computer system we built. The system variants differed in the realization of focus on form (implicit/FonF vs. explicit/FonFS) and the interaction mode (free vs. constrained production). The structure and scenario were selected in such way that the communicative task is realistic and useful for the learner, with the target form either obligatory or natural to employ, thus inducing its incidental use.

2.1. The target form and the task

Form: German dative in prepositional phrases The dative case in German is required as an object of certain spatial prepositions.[2] The case is marked morphologically on the gender-specific article as well as on adjectives and in specific cases on the head noun. Prepositions requiring dative include, among others, *vor* ('in front of'), *hinter* ('behind'), or *zwischen* 'between'. Most locative prepositions used to describe static spatial relations require dative, as does the directional preposition *(bis) zu* ('to'). These prepositions can be elicited in a task which requires spatial descriptions and directions.

Task: Giving directions We designed a directions giving task in such way that an implicit FonF strategy (see below) can most efficiently elicit the forms of interest: The learners were presented with a simplified map of a fictitious campus, with build-

[1] See [10] for the results on the former.

[2] Certain verbs also govern the dative and the dative typically marks the indirect object, however, we do not address these uses here. Note also that the accusative is not likely to be used in the setup we created.

ings, other landmarks and a route to describe. The map was designed in such way that there was a landmark at each point of turn and at the target point. The landmarks were balanced as to gender and the gender was provided on the map to factor out mistakes due to lack of knowledge of the gender. The route included two points of direction change at two different landmarks and the target was placed close to two other landmarks. This setup attempted to elicit the use of dative to describe the point of direction change, in order for the directions to be clear, and made it possible to explicitly induce the use of dative by means of mock-up clarification questions when the learner did not supply them (example below). The choice of landmarks was such that all the genders were represented.

The learners' instructions stated that they were stopped on campus and asked for directions; an explicit request to use the provided map and to describe the indicated route was included. The task description did not, however, contain any hints on using prepositional phrases or paying attention to the dative case.

2.2. Instruction types and language production modes

Two variants of type-written activities were implemented: The *FonF* variant allowed learners to formulate their dialogue contributions freely while providing *implicit corrective feedback on forms* in case the learner made mistakes and *unobtrusively eliciting the target forms* in case the learner did not supply them. The system controlled the interaction by means of a state-based dialogue model. In the *FonFS* activity, learners were asked to produce only the target forms and the feedback *explicitly* informed whether the supplied form was correct.

Implicit FonF/Free production (FonF/FP) In the free-production system variant, the learners were able to type their utterances freely without any restrictions on the language used.[3]

The system classified the learner's input into one of three categories ("TF": "target form"): TF-realized-correct, TF-realized-incorrect, TF-not-realized. It provided implicit feedback in case of errors in the TF by reformulating (*recasting*) the learner's utterance (or parts thereof) in a form of an implicit confirmation type of grounding move (e.g. **L**earner: Hinter das$_{\text{TF-realized-incorrect}}$ Cafe nach links./'Turn left, past the coffee-shop.' **S**ystem: Okay, [hinter dem Cafe nach links]$_{\text{RECAST}}$/'Okay, left past the coffee-shop.') If the target form was not realized, the system tried to elicit it once by asking a clarification question (e.g. **L**: und dann nach links./'and then left.' **S**: [wo soll ich links?]$_{\text{ELICIT}}$/'where do I turn left?')

The system did not correct any other structures except those in focus. In case a learner would give a complete route description in one turn at the start of the dialogue, the system prolonged the interaction by asking to slow down, confirming only the first part of the description, and prompting for continuation.

Explicit FonFS/Constrained production (FonFS/CP) In the constrained production variant the learners' production was restricted to supplying the target form by filling gaps in prescripted dialogue turns as in the example below:

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'

[3]The system implemented two input interpretation strategies: one based on a grammar with mal-rules, and a fall-back strategy based on fuzzy keyword matching. A simple grammar analogous to those employed in spoken dialogue systems (encoded in Java Speech Grammar Format) was created. A parser was extracted from the open-source CMU Sphinx system; `http://cmusphinx.sourceforge.net`.

The learners were allowed three attempts to produce the correct form. The system explicitly signalled incorrect forms with a message and subtracted one point from the learner's "score"; correct forms increased the score by one. After the third unsuccessful attempt the correct utterance was appended to the dialogue and the system generated its next turn.[4]

3. THE EXPERIMENT

Design and procedure The study we conducted used a quasi-experimental design involving students from two German language classes taught by different teachers. Each class was split randomly into two sub-groups. One of the sub-groups was assigned to the free production, **FonF/FP**, condition and the other to the constrained production, **FonFS/CP**, condition.

The groups participated in two sessions with the systems, with one week's break between the sessions. Each session consisted of at least two repetitions of the dialogue activity in different configurations of the map; see Section 2.1.[5] Directly before and after the sessions, the subjects completed grammar tests and performed an analogous directions giving task in a spoken dialogue with a peer (see below). A pre-test was administered before the first session. Immediate post-tests followed the first and the second session. A delayed post-test followed after five weeks' break. After the second session learners completed a biographical/system usability survey.

In order to elicit spoken production data, the participants were asked to work in pairs. They were given landmaps (adapted from Map Task [11]) with five different landmarks including the vocabulary and gender information. One partner's map contained a route to describe, while the other's task was to draw the route. The conversations were recorded upon consent and the data was used to evaluate oral fluency. The recordings were edited to remove irrelevant, non-task-related conversation and to split the recordings into separate files for each participant, the instructions giver, at a given test time. Apart from short confirmations and clarification questions, longer utterances by the other partner were excluded. The resulting samples were between 30 and 90 seconds long. The spoken interaction task was performed in the beginning of each of the two treatment sessions (T1, T2) and directly before the delayed post-test (T3).[6]

Participants Participans of the study were 22 students registered in general German courses at a university. They came from different language backgrounds, were both male and female, with an average age of 24 years, and have been learning German for an average of two years prior to the experiment. Their German proficiency was classified as ranging from A2 to B1+ CEF level [12], based on an initial course placement test. The courses met twice a week for 90 minutes. The experiment took place 6 weeks (approximately 15 instruction hours) into the course. Due to participant drop out and data loss, we could only use data of 13 participants to evaluate oral fluency.

[4]The two system variants were built on the same architecture. The CP variant used a simpler method to map the input to the expected answer (case-insensitive string matching) and a simplified dialogue model because elicitation subdialogues and recasts were not employed.

[5]For the second round, the basic map was kept, but the start point, the goal, the route, and the set of landmarks were changed. The participants took from 5 to 24 minutes (13 mins on average) to complete the activity in the FonF/FP and from ca. 2 and 10 minutes (4:30 average) in the FonFS/CP condition.

[6]We could not conduct the oral elicitation tasks exactly in parallel with the other tests for logistic reasons, thus we only have three samples for each participant instead of four.

4. ANALYSIS AND RESULTS

4.1. Measures of fluency

In evaluating the fluency of learners' oral production we pursued two methods: subjective holistic rating of fluency as perceived by human raters and an analysis of objective temporal measures of the recorded speech data.

Holistic rating of perceived fluency Following a methodology similar to [13] we asked three raters to rank the participants' audio samples. The data were presented using the Rating Test tool[7] which allows raters to click on audio buttons to listen to samples and then to drag the buttons on a two-dimensional pane to form a ranking (i.e. the scale is continuous.)

Procedure Raters were presented with the three (or two if only two recordings existed) speech samples of one participant recorded at different test times and were asked to rank them. The rating instructions were as follows: *How good is the German of the speakers? Can they express themselves clearly and effectively? How fluent are they?* In order to prevent confounding effects raters were explicitly asked to disregard pronunciation and grammatical accuracy since we did not expect the treatment to have an effect on pronuncation and we want to evaluate grammatical accuracy separately in detail, based on transcripts. The samples were randomized: both the participants and the given participants' samples. The was no restriciton on the number of times the samples could be played.

Raters Three raters with background in teaching German as a Foreign Language (GFL) and with differing amount of experience in judging learners' performance were hired for the rating task. Two were experienced teachers, while the third was a fourth year GFL student. After the first round of rating we tested inter-rater agreement (Kendall's coefficient of concordance) and selected a data subset which contained (1) samples rated with low consistency ($W < 0.5$) and (2) a small subset of samples rated with higher consistency. These were re-rated by the same raters in order to assess the intra-rater agreement (consistency).

Agreement In general inter-rater agreement was low ($W = 0.39$ averaged over all samples). Also, intra-rater agreement for two of the three raters was low ($W_1 = 0.68$ and $W_2 = 0.66$, averaged over all samples, versus $W_3 = 0.88$ of the most consistent rater) suggesting that the task is not easy. This impression was also confirmed by the raters themselves. In the results section (below) we report only ratings of the most consistent rater. Spearman's rank correlation coefficient was used to assess the association between rating and test time.

Temporal measures In order to examine temporal correlates of fluency, speech data were transcribed and annotated. The following measures shown to correlate with holistic perception of fluency [14] were extracted: speech rate calculated in syllables per second (**SR-s**) and in words per second (**SR-w**), mean length of pauses calculated as the average length of pauses longer than 0.25 seconds (**MLOP**), mean length of runs calculated as the average number of syllables produced between pauses longer than 0.25 seconds (MLOR; since it is not clear how to treat filled pauses, i.e. nonlexical voiced fillers such as "uh", "um", we calculated two measures: mean runs including syllables with filled pauses, **MLOR-fp**, and without, **MLOR**), and phonation-time ratio (PTR) calculated as proportion of time spent speaking in the time taken to produce the speech sample [15]; three PTR measures were calculated: **PTR-nfp** disregards filled pauses altogether, while **PTR-fp:s** and **PTR-fp:ns** count them as speaking time and non-speaking time, respectively.

Because parametric assumptions were not met, we performed non-parametric analyses: The Friedman test was used for within-subject differences, followed by pairwise post-hoc comparisons using the Wilcoxon signed rank test on groups for which the Friedman test was statistically significant. For between-group comparisons the Mann-Whitney-U test was used. The significance level was set at 0.05, however we also report results at $\alpha = 0.10$ to indicate interesting tendencies.

Table 1: Correlations between test time and rank

Time	FonF/FP (n=6)	FonFS/CP (n=7)	Fisher's r-to-z p-value
T1-T2-T3	0.56	0.34	0.36
T1-T2	0.25	-0.22	0.19
T1-T3	0.55	0.62	0.81
T2-T3	1.00	0.51	< 0.001

4.2. Results

Table 1 presents Spearman's rank correlation coefficient between test time and rank (holistic rating of fluency) for both conditions and all subsets of test times. The p-values in the rightmost column are the significance levels of Fisher's r-to-Z transformation which was used to compare the correlations.[8] The only significant difference was found between the second and third oral test, with a stronger correlation between the rank and test time in the implicit FonF/FP group.

Table 2 shows the means and standard deviations (in parentheses) of all the temporal measures for both experiment conditions, averaged over participants.

The first thing to notice is that at the pre-test the two groups only differed significantly on the mean length of pauses, MLOP, but not on any other measures, with the FonF/FP group showing significantly longer pauses at pre-test than the FonFS/CP group. Since the length of pauses negatively influences the perceived fluency, the FonF group could have been considered less fluent than the FonFS group before treatment. MLOP increased in the FonFS/CP group and decreased the FonF/FP group, however, these changes are not statistically significant.

No significant differences were found in terms of speech rate measured in syllables per second, SR-s. However, for speech rate measured in words per second, SR-w, the FonFS/CP group showed a significant increase between T1 and T3.

On the mean length of runs with filled pauses, MLOR-fp, both groups started off the same and showed no change. Both groups exhibited the same pattern on both MLOR measures: decrease from T1 to T2 and increase from T2 to T3. With filled pauses excluded, MLOR, in the FonFS/CP group these differences are significant. In the FonF/FP group only the decrease between T1 and T3 is significant at $\alpha = 0.10$.

The groups did not show any difference on the PTR-fpns measure (filled pauses not considered as speech). However, on PTR-fps (filled pauses as speech) the FonF/FP group increased between T1 and T3 (at $\alpha = 0.10$), while the FonFS/CP group increased between T1 and T2 ($\alpha = 0.10$). With filled pauses excluded, PTR-nfp, the FonF/FP group increased between T1 and T3 ($\alpha = 0.10$) and at T2 the FonFS/CP group showed higher proportion of phonation time than the FonF/FP group.

[7] http://ratingtest.sourceforge.net

[8] Fisher's transformation is not defined for $r = 1$; 0.99999 was used

Table 2: Means and standard deviations (in parentheses) of temporal measures. Statistically significant differences are marked in bold: **G** indicates between-group difference at the given time, **T** between-time difference within the given group; * $p <= 0.10$, ** $p <= 0.05$

Measure	FonF/FP (n=6)						FonFS/CP (n=7)					
	T1		T2		T3		T1		T2		T3	
SR-s	2.05	(0.30)	1.81	(0.58)	2.19	(0.50)	2.02	(0.66)	2.05	(0.53)	2.31	(0.60)
SR-w	1.29	(0.25)	1.25	(0.35)	1.43	(0.33)	**1.42**$^{\mathbf{T}**}$	(0.41)	1.50	(0.34)	**1.62**$^{\mathbf{T}**}$	(0.36)
MLOP	**0.85**$^{\mathbf{G}**}$	(0.15)	0.80	(0.25)	0.63	(0.13)	**0.62**$^{\mathbf{G}**}$	(0.19)	0.68	(0.25)	0.69	(0.26)
MLOR-fp	**4.76**$^{\mathbf{T}*}$	(1.09)	3.57	(1.32)	**4.19**$^{\mathbf{T}*}$	(0.91)	4.77	(1.89)	4.02	(0.84)	5.43	(2.76)
MLOR	5.37	(1.31)	4.45	(1.81)	4.96	(1.52)	**5.01**$^{\mathbf{T}**1}$	(1.56)	**3.96**$^{\mathbf{T}**1}$	(0.93)	**6.18**$^{\mathbf{T}**1}$	(3.07)
PTR-nfp	**0.67**$^{\mathbf{T}*}$	(0.06)	**0.64**$^{\mathbf{G}*}$	(0.13)	**0.76**$^{\mathbf{T}*}$	(0.10)	0.70	(0.18)	**0.73**$^{\mathbf{G}*}$	(0.09)	0.79	(0.12)
PTR-fps	**0.57**$^{\mathbf{T}*}$	(0.05)	0.56	(0.15	**0.68**$^{\mathbf{T}*}$	(0.10)	**0.60**$^{\mathbf{T}*}$	(0.19)	**0.65**$^{\mathbf{T}*}$	(0.10)	0.70	(0.17)
PTR-fpns	0.72	(0.07)	0.69	(0.10)	0.78	(0.09)	0.76	(0.14)	0.76	(0.07)	0.83	(0.09)

[1] $T1 > T2, T2 < T3$

5. CONLUSION

The presented study investigated the effect of *type-written* computer-based communicative language instruction, implicit FonF and explicit FonFS, on *oral* production. The effect was evaluated using ratings of perceived fluency and temporal measures of speech production. Since the number of participants whose data we were able to analyse was small, it is hard to draw firm conclusions. However, certain tendencies can be observed:

Correlations between holistic ratings and test times are in general low. The only reliable correlation was found between T2 and T3, that is, preceivable improvement took place only after the second session. The apparent difficulty in rating consistently suggests that between-samples differences might have been too subtle. Another problem might be insufficient sample length or lack of explicit training in this type of rating task.

Regarding the temporal measures related to fluency, it is interesting to note that significant improvements show between T1 and T3. In fact, the FonF/FP group shows improvements only at this interval. This suggests that the implicit FonF instruction might have more long-term than intermediate effect. Since the length of runs has been shown to be positively correlated with fluency, the decrease in MLOR between T1 and T3 in the FonF/FP group indicates a decrease in fluency, at least in this respect. However, lower MLOR might also indicate an increase in efficiency, instructions being more concise. Although the FonF/FP group improves on one phonation-time measure, PTR-nfp, between T1 and T3, its performance at T2 is inferior to the other group's. Still, an improvement trend in this group can be observed in terms of reaching the performance of the other group on mean length of pauses after an inferior pre-test.

In summary, the results suggest that the effect of the two types of instruction on fluency are subtle and similar. In fact, the only between-group difference which can be attributed to treatment is in the phonation time excluding filled pauses, PTR-nfp, at T2. We plan to conduct another longer-term experiment with a larger number of participants in order to obtain more reliable statistical results. We also plan to look into the correlation between the learners' performance in the course of the interaction with the system and the spoken fluency as presented here.

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. 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 project ALLEGRO (Project No.: 67 SMLW 1 1 137).

7. References

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

[2] M. H. Long, "Focus on form: A design feature in language teaching methodology." in *Foreign language research in cross-cultural perspective*. Amsterdam: John Benjamins, 1991, pp. 39–52.

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

[4] N. Spada, "Form-focussed instruction and second language acquisition: A review of classroom and laboratory research," *Language Teaching*, vol. 30, no. 02, pp. 73–87, 1997.

[5] A. Poole, "Focus on form instruction: foundations, applications, and criticisms," *The Reading Matrix*, vol. 5, no. 1, pp. 47–56, 2005.

[6] C. Wang and S. Seneff, "A spoken translation game for second language learning," in *Proceedings of AIED-07*, 2007.

[7] W. L. Johnson and S. Wu, "Assessing aptitude for learning with a serious game for foreign language and culture," in *Intelligent Tutoring Systems*. Springer Berlin / Heidelberg, 2008.

[8] J. van Doremalen, H. Strik, and C. Cucchiarini, "Optimizing non-native speech recognition for CALL applications," in *Proceedings of INTERSPEECH-09*, 2009, pp. 592–595.

[9] S.-Y. Yoon, L. Chen, , and K. Zechner, "Predicting word accuracy for the automatic speech recognition of non-native speech," in *Proceedings of INTERSPEECH-10*, 2010.

[10] M. Wolska and S. Wilske, "Form-focused task-oriented dialogues for computer assisted language learning: A pilot study on German dative," in *Proceedings of the SLaTE-2010 Workshop*, 2010.

[11] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[12] J. Trim, B. North, and D. Coste, *Gemeinsamer europïscher Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Niveau A1, A2, B1, B2*. Langenscheidt, 2001.

[13] S. Gass, A. Mackey, M. J. Alvarez-Torres, and M. Fernndez-Garca, "The effects of task repetition on linguistic output," *Language Learning*, vol. 49, pp. 549 – 581, 1999.

[14] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[15] R. Towell, R. Hawkins, and N. Bazergui, "The development of fluency in advanced learners of french," *Applied Linguistics*, vol. 17, no. 1, pp. 84–119, 1996.