Evaluating Task Success in a Dialogue System for Indoor Navigation

**Nina Dethlefs, Heriberto Cuayáhuitl, Kai-Florian Richter,
Elena Andonova, John Bateman**
University of Bremen, Germany
dethlefs@uni-bremen.de

Abstract

In this paper we address the assessment of dialogue systems for indoor wayfinding. Based on the PARADISE evaluation framework we propose and evaluate several task success metrics for such a purpose. According to correlation and multiple linear regression analyses, we found that task success metrics that penalise difficulty in wayfinding are more informative of system performance than a success/failure binary task success metric.

1 Introduction

Wayfinding in (partially) unknown environments poses a considerable challenge for humans. Our work addresses indoor navigation within complex buildings that present significant navigational challenges to new and infrequent visitors. This application scenario is of increasing relevance nowadays, as building complexes become larger and greater attention is paid to making them accessible for a broader range of users. Despite this, automatic systems generating natural language-based in-advance route descriptions have received little attention to date. Previous work consists of either attempts to involve primarily visual support or the quality and effects of the language component are taken for granted (Kray et al., 2005; Callaway, 2007; Kruijff et al., 2007). In particular, there appears to be a lack of agreed-on evaluation metrics for assessing the performance of dialogue systems in the wayfinding domain. This paper addresses this lack and suggests to employ metrics that are sensitive to task difficulty for evaluation of such systems. We present a study that confirms the usefulness of our proposed metrics.

2 System Architecture

We base our study on data collected from a dialogue system that assists users in indoor wayfinding using text-based natural language input and output. The system architecture consists of four different modules for the tasks of Natural Language Understanding, dialogue management, route instruction generation and Natural Language Generation. The following is a sample dialogue translated from German (S=System, U=User):

```
S: Do you need a route description on this
    level? Just write your question into
    the text field.
U: Where can I find room number 3180?
S: I have found the following rooms: b3180,
   a3180. Which are you looking for?
U: b3180
S: Please turn around and go straight until
   the next hallway on the left-hand side.
   Turn left and go until the door b3180
   on the right-hand side.
   Do you have further questions?
```

Since the task success metrics proposed in this paper aim to be independent of specific architectural decisions, we refer the reader to Cuayáhuitl et al. (2010) for details on our indoor navigation dialogue system.

3 Experimental Setting

3.1 Evaluation methodology

Evaluation of the system was performed using objective and subjective metrics mostly derived from the PARADISE framework (Walker et al., 2000). We used the following quantitative metrics. First, the group of *dialogue efficiency* metrics includes ‘system turns’, ‘user turns’, and ‘elapsed time’ (in seconds). The latter includes the time used by both conversants, from the first user utterance until the last system utterance. Second, the group of *dialogue quality* metrics consists of percentages of parsed sentences, sentences with spotted keywords, and unparsed sentences. Third, the

Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue

group of *task success* metrics includes the well known success/failure Binary Task Success (BTS) defined as

$$\text{BTS} = \begin{cases} 1 & \text{for} \quad \text{Finding the Target Location (FTL),} \\ & \quad\quad\;\; \text{with or without problems} \\ 0 & \quad\quad\;\; \text{otherwise.} \end{cases}$$

Because this metric does not penalise difficulty in wayfinding, we propose and evaluate the following metrics — referred to as Graded Task Success (GTS) — that penalise with different values:

$$\text{GTS}^a = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 0 & \quad\quad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^b = \begin{cases} 1 & \text{for} \quad \text{FTL with none or small problems} \\ 0 & \quad\quad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^c = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 1/2 & \text{for} \quad \text{FTL with small problems} \\ 0 & \quad\quad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^d = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 2/3 & \text{for} \quad \text{FTL with small problems} \\ 1/3 & \text{for} \quad \text{FTL with severe problems} \\ 0 & \quad\quad\;\; \text{otherwise.} \end{cases}$$

We coded difficulty in wayfinding, using the categories 'no problems', 'small problems' and 'severe problems' as follows. The value of 1 was given when the user finds the target location without hesitation, the value with 'small problems' was given when the user finds the location with slight confusion(s), and the value with 'severe problems' was given when the user gets lost but eventually finds the target location. The motivation behind using task success metrics that penalise differently the difficulty in wayfinding was to discover a metric that correlates highly with user satisfaction. Such a metric aims to be more informative for assessing task success performance than the traditional binary task success metrics. We tried four different graded metrics, GTS^a - GTS^d, in order to find the metric that best predicted user satisfaction. For the qualitative evaluation we used the subjective metrics described in (Walker et al., 2000).

3.2 Evaluation setup

Twenty-six native speakers of German participated in our study with an average age of 22.5 and a gender distribution of 16 female (62%) and 10 male (38%). Each subject received six dialogue tasks, corresponding to locations to find, which resulted in a total of 156 dialogues. Dialogues consisted of differing numbers of High-Level instructions (HLIs). High-Level Instructions (HLIs) encapsulate a set of low-level instructions (e.g., 'go straight', 'turn left', 'turn around') and are based on major direction changes. Two dialogue tasks used 2 High-Level Instructions (HLIs) such as those shown in the dialogue on page 1. Two other tasks used 3 HLIs, and two used 4 HLIs. The tasks were executed pseudorandomly (from a uniform distribution), so that the order of task execution would not impact on the user ratings. The participants were asked to request a route from the system using natural language, optionally take notes, and then follow the system instructions closely trying to find the locations. They were not allowed to ask anybody for help. Participants could give up when they were unable to find the target location by telling that to the assistant that followed them. It was the task of this assistant as well to judge and take note of the difficulties that subjects encountered in their wayfinding task as described in the previous section. At the end of each dialogue, participants were asked to fill in a questionnaire for obtaining qualitative results using a 5-point Likert scale, where 5 represents the highest score.

4 Experimental Results

Table 1 summarises our results for the quantitative and qualitative metrics. It can be observed from the dialogue efficiency metrics (first group) that user-machine interactions involved short dialogues in terms of turns and interaction time. Once users received instructions from the system, they tended not to ask further. With regard to dialogue quality (second group), we noted that our grammars need to be extended in coverage and that the keyword spotter proved vital in the dialogues. The analysis of task success measures (third group) revealed very high binary task success, and lower scores for the other task success metrics.

4.1 Correlation analysis

In a correlation analysis between task success measures and user satisfaction we obtained the results displayed in Table 2. This can be interpreted as follows: while all metrics correlate moderately with overall user satisfaction, the metrics taking task difficulty into account correlate higher. A more detailed analysis of the corre-

Table 1: Mean values of our evaluation metrics for our wayfinding system based on 156 dialogues, organised in four groups: dialogue efficiency, dialogue quality, task success and user satisfaction.

Measure	Score
System Turns	2.30 ± 0.3
User Turns	1.52 ± 0.5
System Words per Turn	41.30 ± 4.0
User Words per Turn	4.79 ± 2.1
Interaction Time (secs.)	22.14 ± 18.4
Session Duration (secs.)	2014.62 ± 393.2
Parsed Sentences (%)	16.7 ± 16.0
Spotted Keywords (%)	79.9 ± 17.0
Unparsed Sentences (%)	3.4 ± 0.5
Binary Task Success (%)	94.9 ± 8.3
Graded Task Success^a (%)	71.4 ± 15.0
Graded Task Success^b (%)	87.8 ± 15.0
Graded Task Success^c (%)	81.4 ± 13.3
Graded Task Success^d (%)	87.6 ± 8.3
(Q1) Easy to Understand	4.46 ± 0.8
(Q2) System Understood	4.65 ± 0.8
(Q3) Task Easy	4.29 ± 0.9
(Q4) Interaction Pace	4.63 ± 0.5
(Q5) What to Say	4.66 ± 0.7
(Q6) System Response	4.56 ± 0.6
(Q7) Expected Behaviour	4.45 ± 0.8
(Q8) Future Use	4.31 ± 0.9
Overall User Satisfaction (%)	90.0 ± 7.3

lation between task success metrics and individual user satisfaction metrics revealed the following. First, the binary task success showed lower correlations than the other metrics in the subjective metric 'easy to understand' (Q1). Second, while there is no correlation between the subjective metric 'future use' (Q8) and binary task success, the other metrics reveal a moderate correlation. Third, while binary task success shows a moderate correlation for 'task easy' (Q3), the other metrics show a high correlation. Therefore, we can conclude that the task success metrics that penalise difficulty in wayfinding are more informative of user-system interaction performance for indoor wayfinding than the BTS metric. Furthermore, there was no correlation between the number of high-level instructions and overall user satisfaction, i.e. user satisfaction was independent of instruction length (our system performed equally well for short and long routes).

Table 2: Correlation coefficients between task success and user satisfaction measures (significant at $p < 0.05$).

Measure	BTS	GTS^a	GTS^b	GTS^c	GTS^d
Q1	.47	.44	.54	.49	.54
Q2	.20	.17	.19	.19	.20
Q3	.53	.67	.71	.71	.76
Q4	.21	.26	.24	.24	.28
Q5	.20	n.s.	.17	.18	.18
Q6	n.s.	n.s.	n.s.	n.s.	n.s.
Q7	.31	.35	.44	.40	.44
Q8	n.s.	.39	.32	.40	.39
Overall	.43	.52	.55	.55	.60

Note: n.s. - not significant.

4.2 Multiple linear regression analysis

In order to identify the relative contribution that different factors have on the variance found in user satisfaction scores, we performed a standard multiple linear regression analysis on our data. According to the PARADISE framework (Walker et al., 1997), performance can be modeled as a weighted function of task-success measure and dialogue-based cost measures. The latter represent the measures summarised under dialogue efficiency and dialogue quality above. We normalised all task success and cost values to account for the fact that they can be measured on different scales (seconds, percentages, sum, etc.), according to $\mathcal{N}(x) = \frac{x-\bar{x}}{\sigma_x}$, where σ_x corresponds to the standard deviation of x. Then we performed several regression analyses involving these data.

Results revealed that the metrics 'user turns' and 'task success' (for GTS^a, GTS^c and GTS^d) were the only predictors of user satisfaction at $p < 0.05$. The other task success measures were not significant (with BTS at $p = 0.39$ and GTS^b at $p = 0.17$). These results confirm our claim that task success metrics that consider difficulty in wayfinding (specifically GTS^a, GTS^c and GTS^d) are more informative with respect to user satisfaction in the wayfinding domain than a binary success/failure metric. Subjects seem to be sensible to problems they encounter in their wayfinding tasks, which are expressed in their ratings of the system.

4.3 Estimation of a performance function

We use the following equation to obtain a performance function (Walker et al., 1997):

$$\text{Performance} = (\alpha * \mathcal{N}(k)) - \sum_{i=1}^{n} \omega_i * \mathcal{N}(c_i),$$

where, α is a weight on the task success metric k (to be replaced by any of our proposed metrics), and ω_i is a weight on the cost functions c_i. $\mathcal{N}$ represents the normalised value of c_i. Based on the results of our first regression analysis, we ran a second analysis using those variables that were significant predictors in the first regression, i.e. the number of user turns and task success metrics GTS^a, GTS^c and GTS^d. We analysed the correlation between these variables, which resulted in weak negative correlations. We obtained the following performance function for task success metrics GTS^c and GTS^d (because those two accounted for most of the variance in user satisfaction), where UT refers to 'User turns':

$$\text{Performance} = 0.38\mathcal{N}(GTS^{c,d}) - 0.87\mathcal{N}(UT),$$

suggesting that the more successful and efficient the interaction, the better. These results show that GTS^c, GTS^d and UT are significant at $p < 0.01$, and the combination of UT and each of GTS^c and GTS^d account for 62% of variance in user satisfaction. This performance function can be used in future evaluations of the system.

5 Discussion

The idea of taking different degrees of task difficulty into consideration in evaluation is not entirely new (Tullis and Albert, 2008). However, to the best of our knowledge, there have been no previous studies that demonstrated that these metrics do indeed show a higher correlation with user satisfaction scores than the BTS metric, which is typically used to assess task success. This finding was supported by an evaluation in a real environment using an end-to-end dialogue system, and was based on PARADISE, a generic framework for the evaluation of (spoken) dialogue systems. The proposed metrics can therefore be regarded as a useful and important step contributing to the understanding of the performance of situated dialogue systems. Further, our proposed metrics address the lack of standardised evaluation metrics in the wayfinding domain in particular. We presented a concrete performance function that can help future system development in the domain by allowing the estimation of relative contributions of different task success metrics and cost function towards overall user satisfaction.

6 Conclusion

In this paper we addressed the assessment of dialogue systems for indoor navigation using the PARADISE framework and different task success metrics. We found that task success metrics that take difficulty in wayfinding into account correlate higher with overall user satisfaction than a binary task success metric. In addition, a more detailed correlation analysis for subjective metrics of user satisfaction confirmed that our proposed metrics are more informative of system performance for indoor wayfinding than the binary success/failure metric. This result was confirmed by a multiple linear regression analysis that tested for the relative contribution to variance in user satisfaction of different task success metrics and cost measures. Future work can apply these metrics to dialogue systems with different input and output modalities.

Acknowledgements

Part of this work was supported by the DFG Transregional Collaborative Research Center SFB/TR8 "Spatial Cognition".

References

Charles Callaway. 2007. Non-localized, interactive multimodal direction giving. In I. van der Sluis, M. Theune, E. Reiter, and E. Krahmer, editors, *Proceedings of the Workshop on Multimodal Output Generation MOG 2007*, pages 41–50. Centre for Telematics and Information Technology (CTIT), University of Twente.

Heriberto Cuayáhuitl, Nina Dethlefs, Kai-Florian Richter, Thora Tenbrink, and John Bateman. 2010. A dialogue system for indoor wayfinding using text-based natural language. In *International Journal of Computational Linguistics and Applications, ISSN 0976-0962*.

Christian Kray, G. Kortuem, and A. Krüger. 2005. Adaptive navigation support with public displays. In Robert St. Amant, John Riedl, and Anthony Jameson, editors, *Proceedings of IUI 2005. ACM Press, New York*, pages 326–328.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138, March. Special Issue on Human-Robot Interaction.

Tom Tullis and Bill Albert. 2008. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *ACL*, pages 271–280.

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3):363–377.

Guiding the User when Searching Information on the Web

Marta Gatius **Meritxell González**
Technical University of Catalonia
Barcelona, Spain
{gatius,mgonzalez}@lsi.upc.edu

Abstract

This paper describes how we approach the problem of guiding the user when accessing informational web services. We developed a mixed-initiative dialogue system that provides access to web services in several languages. In order to facilitate the adaptation of the system to new informational web services dialogue and task management were separated and general descriptions of the several tasks involved in the communication process were incorporated.

1Introduction

This paper describes how we approach the problem of guiding the user when accessing informational web services. We designed a dialogue system (DS) for accessing different types of applications in several languages. The results of the evaluation of the first prototype are described in (Gatius and González, 2009). In order to improve both the functionality and adaptability of the DS we have studied the most appropriate representation of the general and application-specific conceptual knowledge involved when helping the user to access informational services.

When providing access to information-seeking applications DSs use an underspecified set of constraints to restrict the search rather than a defined user's goal (which can be broken down into tasks and subtasks). Hence, the main tasks for DSs providing access to an informational service consist of guiding the user to give the needed constraints as well as presenting in an appropriate way the results. There have been several approaches to face this problem (Rieser and Lemon, 2009; Steedman and Petrick, 2007; Varges et al., 2009). Our approach consists of separating completely dialogue management from task management (following other relevant proposals (Allen et al., 2001)), and defining the general tasks involved when accessing informational services. Besides, general mechanisms using the two main knowledge bases of the system (the dialogue context and the domain conceptual knowledge) are used to relax the query constraints and to state additional constraints.

2 Dialogue and Task Management

The DS we developed consists of five independent modules: the language understanding, the dialogue manager (DM), the task manager, the language generator and the user model, used to adapt automatically the dialogue strategies. Additionally, there are two main data structures accessible for all modules: the information state, representing the dialogue context and the conceptual knowledge, describing the application domain.

The DM follows the information state update model, which provides a complete separation of dialogue and task management. The DM uses communication plans to determine the next system actions that could satisfy user's requirements. These plans are generated (semi)-automatically when a new service is incorporated into the DS by adapting the general communication plan for the service type to the particular service specifications.

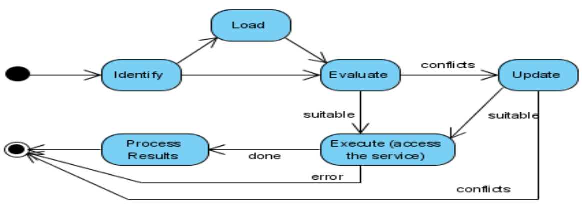

Figure 1: Task Management in the Dialogue System

Figure 1 shows task management in the DS. Main tasks performed by the task manager are the following: identification of the required web service and the specific service task, completion of the data obtained from the user, access to the service and presentation of the results.

Once the communication starts and the first intervention of the user has been interpreted and passed to the task manager, it has to identify the service and the specific service task that has to be accessed. Then, an instantiation of the specific task is generated. There are several general task descriptions for each service type, for the informational services two tasks are considered: find a list of items and describing an item.

Finally, the task manager accesses the web service and decides the most appropriate presentation of the results obtained.

3 Accessing InformationalServices

The tasks involved when the system guides the user to access an informational service are shown in Figure 2. Circled blocks represent the specific information the DM has to obtain from the user: the searched data (requestedData or output parameters) and the data constraining the query (queryConstraints or input parameters). Rectangles represent the three different tasks processing the resulting data: describing a particular item, collecting a list of items and summarizing the results obtained. Colored blocks correspond to the three different processes considered when updating constraints: relaxation, using default values and adding new constraints.

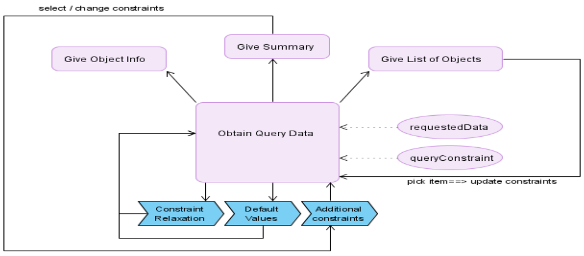

Figure 2: The tasks involved in information-seeking

The process of obtaining the query constraints from the user could be complex, as they are not gathered in a predetermined order. The task manager determines whether a complete query can be generated or if additional information has to be obtained from the user. If the service's definition includes default values, they can be included to complete the query. Parameter values appearing in previous turns can also be used.

The information obtained from the service has to be processed. Four different situations are distinguished: the result is only one item, the number of items obtained belongs to a predefined range, there are too many results and there are no results. In case there is only one item a detailed description of this item is given. In case the number of results is acceptable, a list enumerating all of them is presented to the user, suggesting him to pick up one. In the two latter cases the constraints have to be updated.

In the specific case that there are no results, the task manager can automatically relax the constraints and execute the query again. The constraints can be relaxed at the level of the query and at the level of the parameter's values. In the former, the system removes one or more of the query constraints. In the latter, the system updates the value for one or more of the constraints. The conceptual knowledge base is used to relax the constraints. If taxonomies describing the domain have been incorporated, a class is substituted by the upper class (for example, if the user asks for *drama movies* and there are none, the upper class *movies* would be used). Several strategies for data common to several applications (such as dates and locations) are already considered.

In the specific case that too many items are obtained from the service, the system presents a summary of the results. Information suggesting possible additional constraint values could also be given to the use.

5 Conclusions and Future Work

In order to improve the functionality and adaptability of our DS when guiding the user to accessing informational service we have studied the general and the application specific conceptual knowledge involved in the communication process. In our system this knowledge has been represented as a general scheme from which the communication plans for each informational service are generated and general task that are instantiated for each service. The resulting architecture facilitates the integration of other application types into the system since the task models can be easily extended and adapted.

Future work could include the processing of user's questions which answer involves the processing of data obtained from several web services.

References

J.F. Allen, G. Ferguson and A. Stent 2001. An Architecture for More Realistic Conversational Systems. Intelligent User Interfaces Conference.

M. Gatius and M. González 2009. A Flexible Dialogue System for Enhancing Web Usability. 18th World Wide Web Conference.

M. Steedman and R. Petrick 2007. Planning Dialog Actions. 8th SIGdial Workshop on Discourse and Dialogue.

V. Rieser and O. Lemon 2009. Does this list contain what you were searching for? Learning adaptive dialogue strategies for Interactive Question Answering. Natural Language Engineering, 15(1):57- 375. Cambridge University Press.

S. Varges, F. Weng and H. Ponbarry 2009. Interactive question answering and constraint relaxation in spoken dialogue systems. Natural Language Engineering,15(1).Cambridge University Press.

Semantics and pragmatics of negative polar questions

Maria H. Golka

Chair of Logic and Cognitive Science
Institute of Psychology
Adam Mickiewicz University
Poznań, Poland

Maria.Golka@gmail.com

Abstract

This paper aims to provide a literature review about the meaning and use of negative polar (yes/no) questions and complete it with some Polish data. Semantic and pragmatic factors will be discussed. Attention will be drawn to the fact that most of research concentrate on interrogatives themselves, neglecting their possible answers, whereas the latter may be very informative about the nature of the former.

1 Introduction

From a logical semantic point of view, since a polar question $?\phi$ and its negative counterpart $?\neg\phi$ have the same answers, they are logically equivalent (Groenendijk and Stokhof, 1997). It is obvious, however, that if we consider the natural language use of negative polar interrogatives like for example the one in (2), we can not consider them equivalent to positive ones, as in (1).

(1) Is Jane coming?

(2) Isn't Jane coming?

2 Pragmatic and semantic factors

If negative and positive polar questions are semantically the same, why would we use both of them? Considering some common pragmatic intuitions (captured by numerous concepts like Principle of Economy, Principe of Least Effort, Gricean Maxim of Manner or the minimization of cognitive effort in terms of Relevance Theory) negative interrogatives would not be used if their meaning were not at least pragmatically different from that of positive ones.

These intuitions are confirmed by classic experimental results in psycholinguistics. Syntactic transformations of kernel sentences into other structures like interrogatives or negatives are rather charging for the cognitive system. The syntactic form of a sentence (whether it is an active, passive, interrogative or negative clause) seems to be something distinct and more difficult to recall than its semantic content (Mehler, 1963). Syntactically complex sentences, like questions or negatives, require more capacity of immediate memory. Sentences which are both interrogatives and negatives are the ones that are the most hard to process (Savin and Perchonock, 1965). The usage of negative questions that are semantically equivalent to the positive ones but much more difficult to process can thus be explained by pragmatic factors only.

Nevertheless, some approaches find the nature of the distinction between negative and positive polar questions semantic (e.g. Romero and Han, 2004). They are consistent with Ladd's (1981) observations. As Ladd points out, negative polar questions are systematically ambiguous: in case of the "outside negation" reading the speaker believes that the proposition under question is true, whereas in the "inside negation" one the speaker believes it is false.

In this paper we will discuss some examples which show that in Polish Ladd's ambiguity is much more difficult to capture. We will also take into account the possible answers to questions of this kind. It has not been done by most of authors, but it turns out that if we consider the dialogic factors (which in case of questions seem to be very important), the nature of negative vs. positive polar questions distinction appears to be pragmatic. We will argue that even if the internal ambiguity of negative polar questions is due to semantic factors, it is still likely that the distinction between positive and negative questions is pragmatic.

3 Ladd's ambiguity in Polish

Most of the papers on the subject of Ladd's ambiguity (e.g. Romero and Han, 2004; Reese, 2006) discuss polar questions with preposed negation (English interrogative sentences with a negated auxiliary verb) as the one in (2) and ex-

Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue

clude from consideration interrogative sentences with non-preposed negation, as the one in (3) which permit a neutral interpretation in an unbiased context.

(3) Is Jane not coming?

Since in Polish polar interrogatives are formed by means of an interrogative particle *czy* or with intonation alone, the distinction like that between (2) and (3) is nonexistent. Instead, we have only one type of structure which is rather similar to the structure of an affirmative clause and can by preceded (4) or not (5) with the interrogative particle. This structure conveys all the three readings discussed in the literature (Ladd's outside and inside negation readings, and the neutral one).

(4) Czy Jane nie przychodzi?
INTERR. PART. Jane NEG $\text{come}_{\text{3SG, PRES.}}$

(5) Jane nie przychodzi?
Jane NEG $\text{come}_{\text{3SG, PRES.}}$

In Polish, the word order within a sentence is much less strict than the one in English. Consequently, a Polish equivalent of an ambiguous negative polar interrogative, like (6) (the example of Ladd, 1981) would be more naturally represented by a pair of sentences with different word orders where (7a) expresses the outside negation reading, whereas (8a) the inside negation one.

(6) Isn't there a vegetarian restaurant around here?

(7a) Nie ma w okolicy wegetariańskiej restauracji?
NEG $\text{be}_{\text{3SG, PRES}}$ in $\text{neighborhood}_{\text{LOC, SG}}$ $\text{vegetarian}_{\text{GEN, SG}}$ $\text{restaurant}_{\text{GEN, SG}}$

(8a) Nie ma wegetariańskiej restauracji w okolicy?
NEG $\text{be}_{\text{3SG, PRES}}$ $\text{vegetarian}_{\text{GEN, SG}}$ $\text{restaurant}_{\text{GEN, SG}}$ in $\text{neighborhood}_{\text{LOC, SG}}$

This difference in word order seems to corroborate Reese's (2006) intuition that "there is no semantic (…) difference between "outside" and "inside" negation. Rather, what is at issue is whether negation targets the core meaning of an utterance or some secondary meaning".

Further inspection reveals some problems with the inside negation reading of interrogatives constructed with the particle *czy*. Interrogatives like (7b) and (8b) are acceptable but none of them can convey an inside negation reading. It seems that the presence of *czy* can somehow trigger the outside negation or neutral understanding of negative polar questions.

(7b) Czy nie ma w okolicy wegetariańskiej restauracji?

(8b) Czy nie ma wegetariańskiej restauracji w okolicy?

Another very interesting phenomenon is the use of the particle c*zyż*. This form is used to construct rhetorical questions and simultaneously deny the proposition under question. Hence, the negative question preceded with *czyż* conveys an affirmative assertion. This kind of construction seems to be a paradigmatic example of an outside negation interrogative.

4 Conclusions

As we have seen, the origins of negative polar questions are hard to define. There is some evidence suggesting that their nature is pragmatic as well as some other evidence, showing their semantic nature. In this paper we try to bring together these two approaches. We provide some evidence from Polish language, as well as evidence about answers. Thus a mixed, semantic-pragmatic model is needed to describe the meaning and use of negative polar interrogatives..

References

Groenendijk, Jeroen and Martin Stokhof. 1997. Questions. In van Benthem, J. and A. ter Meulen, editors, *Handbook of Logic and Language*. Elsevier, Amsterdam, pages 1055-1124.

Ladd, Robert D. 1981. A First Look at the Semantics and Pragmatics od Negative Questions and Tag Questions. In *Proceedings of Chicago Linguistic Society*, pages 164-171.

Mehler, Jacques. 1963. Some effects of grammatical transformations on the recall of English sentence. *Journal of Verbal Learning and Verbal Behavior,* 2(4): 346-351.

Reese, Brian J. 2006. The Meaning and Use of Negative Polar Interrogatives. In Bonami, O. and P. Cabredo Hofherr, editors, *Empirical Issues in Syntax and Semantics 6*, pages 331-354.

Romero, Maribel & Chung-Hye Han. 2004. On Negative Yes/No Questions. *Linguistics and Philosophy,* 27(5): 609-658.

Savin, Harris B. and Ellen Perchonock. 1965. Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, 4(5): 348-353.