Empirical Studies for Intuitive Interaction

Iryna Gurevych* and Robert Porzel

European Media Laboratory, GmbH
Schloss-Wolfsbrunnenweg 33
D-69118 Heidelberg, Germany
firstname.lastname@eml-d.villa-bosch.de
* Current affiliation EML Research, gGmbH

Summary. We present three types of data collections and their experimental paradigms. The resulting data was employed to conduct a number of annotation experiments, create evaluation *gold standards* and train statistical models. The data, experiments and their analyses highlight the importance of data-driven empirical laboratory and field work for research on intuitive multi-modal human-computer interfaces.

1 Introduction

Research on dialogue systems in the past has focused on engineering the various processing stages involved in dialogical human-computer interaction (HCI) - e. g., robust automatic speech recognition, intention recognition, natural language generation or speech synthesis (cf. [1, 4, 2]). Development of such dialogue technologies involves experimental work.

Many issues in the semantics and pragmatics of dialogue can be formulated as empirical topics. Therefore, the analysis of sufficient amounts of data is necessary to study a specific phenomenon or train statistical models. Given a certain hypothesis, the researcher has to find out whether there is empirical evidence for the phenomenon in question, how frequent and important it is. Also, rigorous large-scale evaluations of the algorithms is a driving force of the engineering progress. Such evaluations are only possible if sufficient amounts of annotated data are available. In this case, the output of the computer program (an *answer*) is compared to the gold standard (a *key*) defined by the annotated data [15].

Alongside these efforts the characteristics of computer-directed language have also been examined as a general phenomenon (cf. [20, 18, 5]). In the following sections we describe new paradigms for collecting dialogue data that can be employed for a variety of necessary examinations that improve the performance and reliability of conversational dialogue systems.

2 Collection and Usage of *Hidden Operator Test* Data

The data collected in the first experiment was employed in research on empirical topics, such as semantic post-processing of the speech recognizer's outputs and assigning domains to speech recognition hypotheses. We designed a set of annotation schemata and performed experiments with human subjects resulting in a set of annotated corpora. The annotated data was, then, employed in order to:

- test whether the annotators are able to annotate the data reliably;
- produce a gold standard for automatic evaluations of the algorithms;
- create data models based on human annotations;
- produce training datasets for statistical classifiers.

2.1 Data collection setup

The data collection was conducted following an iterative data collection approach [14]. This approach has been developed within EMBASSI research project[1] and employes the *Hidden Operator Test* paradigm. In this experimental setup, the SmartKom system was simulated by an operator with the help of predefined dialogue scripts (see Figure 1). 29 subjects were prompted to say certain inputs in 8 dialogues, resulting in 1479 turns (see Figure 2).

Fig. 1. Operator in hidden operator tests.

Fig. 2. Subject in hidden operator tests.

Due to the controlled experimental setting and prompts appearing on a per-turn basis, each user turn in the dialogue corresponded to a single intention, e.g., route request or sights information request. All user turns were recorded in separate audio files. The audio files obtained from the *Hidden Operator Tests* were then sent to two differently configured versions of the

[1] See www.embassi.de.

SmartKom system. The data describing our corpora is given in Table 1.[2] The first and the second system's runs are referred to as *Dataset 1* and *Dataset 2* respectively.

	Dataset 1	*Dataset 2*
Number of dialogues	232	95
Number of utterances	1479	552
Number of SRHs	2.239	1.375
Number of coherent SRHs	1511	867
Number of incoherent SRHs	728	508

Table 1. Descriptive corpus statistics.

The corpora obtained from these experiments were further transformed into a set of annotation files, which can be read into a GUI-based annotation tool MMAX [9, 10]. This tool can be adopted for annotating different levels of information, e.g., semantic coherence and domains of utterances, the best speech recognition hypothesis in the N-best list, as well as domains of individual concepts. The two annotators were trained with the help of an annotation manual. A reconciled version of both annotations resulted in the gold standard. In the following, we describe a set of annotation experiments performed on the basis of the *Hidden Operator Test* data and present their results. Also, we sketch some applications of the annotated corpora for empirically grounded research.

2.2 Annotation experiments

Semantic coherence and domains of SRHs

In this annotation experiment, the task of annotators was to classify the speech recognition hypotheses (SRHs) as either *coherent* or *incoherent*. We defined semantic coherence as well formedness of an SRH on an abstract semantic level [8]. The SRHs from *Dataset 1* were randomly mixed in order to prevent them from being annotated in the discourse context. The resulting Kappa statistics [3] over the annotated data yielded $\kappa = 0.7$. This indicates that human annotators can distinguish between coherent samples and incoherent ones.

In the second step, the coherent SRHs from both *Dataset 1* and *Dataset 2* were labeled with respect to their domains. We restricted the annotation to coherent SRHs only, as automatically assigning domains to SRHs should only be done if the utterance is a coherent one. The annotations resulted in

[2] We also include the distribution of coherent and incoherent speech recognition hypotheses, which was computed on the basis of annotation experiments described in Section 2.2.

two corpora of coherent speech recognition hypotheses labeled for at least one domain category. The percentage of SRHs with one and more domain attributions can be found in Table 2. It shows that the vast majority of hypotheses (ca. 90%) could be unambiguously classified in a single category. The class distribution is given in Table 3. The majority class is the *Route Planning* domain, while some of the categories, e.g., *Personal Assistance*, *Off Talk* occur rarely in the data.

	Dataset 1		*Dataset 2*	
Number of domains	*Annotator 1*	*Annotator 2*	*Annotator 1*	*Annotator 2*
1	90.06%	87.11%	90.77%	88.7%
2	6.94%	11.27%	9.11%	11.19%
3	3.01%	1.28%	0.16%	0%
4	0%	0.35%	0%	0%

Table 2. Multiple domain assignments to a single SRH.

	Dataset 1		*Dataset 2*	
Domain	*Annotator 1*	*Annotator 2*	*Annotator 1*	*Annotator 2*
Electr. Program Guide	14.43%	14.86%	12.13%	11.73%
Interaction Management	15.56%	15.17%	10.02%	9.24%
Cinema Information	5.32%	8.7%	4.01%	3.43%
Personal Assistance	0.31%	0.3%	0%	0%
Route Planning	37.05%	36%	44.2%	46.31%
Sights	12.49%	12.74%	21.94%	21.08%
Home Appliances Control	14.12%	11.22%	7.59%	8.1%
Off Talk	0.72%	1.01%	0.11%	0.1%

Table 3. Class distribution for domain assignments.

Figure 3 presents the Kappa coefficient values computed for individual domain categories in *Dataset 1*.[3] *P(A)* is the percentage of agreement between annotators. *P(E)* is the percentage that we would expect them to agree by chance. The annotations are generally considered to be reliable if $K > 0.8$. This is true for all classes except those which are under-represented in this dataset (cf. Table 3).

Best SRH in the N-best lists

For this study a markable, i.e. an expression to be annotated, is a set of SRHs (N-best list) related to a single user's utterance. The number of markables in

[3]The Kappa coefficient was not computed for *Dataset 2*, but we expect the results to be very similar. For *Dataset 2*, we employed a different agreement measure described in a separate experiment.

	P(A)	*P(E)*	*Kappa*
Electr. Program Guide	0.9743	0.7246	0.9066
Interaction Management	0.9836	0.7107	0.9434
Cinema Information	0.9661	0.8506	0.7229
Personal Assistance	0.9953	0.9930	0.3310
Route Planning	0.9777	0.5119	0.9544
Sights	0.9731	0.7629	0.8865
Home Appliances Control	0.9626	0.7504	0.8501
Off Talk	0.9871	0.9780	0.4145

Fig. 3. Kappa coefficient for separate domains.

Numb. of dom.	*Ann. 1*	*Ann. 2*
0	4.76%	16.45%
1	60.17%	44.16 %
2	22.51%	18.18%
3	5.63%	10.82%
4	1.73%	7.79%
5	2.16%	1.3%
6	0.43%	0.43%
7	0.87%	0%
8	1.73%	0.87%

Fig. 4. Domain assignments to concepts.

Dataset 2 employed here corresponds to the number of utterances 552. The annotators saw the SRHs together with the transcribed user utterances. For each utterance a single SRH had to be labeled as the best one. The guidelines for selecting the best SRH were:

- how well the respective SRH captures the intention contained in the transcribed user's utterance;
- if several SRHs capture the intention equally well, the actual word error rate had to be considered.

The Kappa coefficient - often applied to measure the degree of inter-annotator agreement - is not appropriate in this experiment, as its calculation is class-based. But the number of SRHs underlying the best SRH selection per utterance is varying. Therefore, we computed the percentage of utterances, where the annotators agreed on the correct solution resulting in 95.35% of inter-annotator agreement. This number suggests a rather high degree of reliability for identifying the best SRH by humans.

Best conceptual representation and domains of SRHs

The best *conceptual representation* and the domains of coherent SRHs from *Dataset 2* were annotated. Due to lexical ambiguity of individual words, each SRH can be mapped to a set of possible interpretations called *conceptual representations* (CR). E.g., the German word *kommen* can be mapped either to the ontological concept `MotionDirectedProcess` or to the concept `WatchPerceptualProcess`. The algorithms operating on the ontology convert each utterance to a set of *conceptual representations*. As a consequence, they must be disambiguated automatically within the running system as well as manually for evaluation purposes.

867 SRHs used in this experiment are mapped to 2853 CRs, i.e., on average each SRH is mapped to 3.29 *conceptual representations*. The annotators' agreement on the task of determining the best CR resulted in ca. 88.93%. For the task of domain annotation, we computed two kinds of agreements. The first measure is the absolute agreement $Prec_{mark}$, when the annotators agreed on all domains for a given markable, i.e. an SRH. This resulted in ca. 92.5%. The second measure $Prec_{dom}$ denotes the agreement on individual domain decisions (6936 overall) and amounted to ca. 98.92%.

Domains of ontological concepts

In the last experiment, ontological concepts were annotated with zero or more domain categories.[4] The percentage of ambiguous domain attributions is given in Figure 4. We extracted 231 concepts from the lexicon, which is a subset of ontological concepts occurring in our data. The annotators were given the textual descriptions of all concepts. These definitions are supplied with the ontology. Again, we computed two kinds of inter-annotator agreement. In the first case, we calculated the percentage of concepts, for which the annotators agreed on all domain categories $Prec_{mark}$, resulting in ca. 47.62%. In the second case, the agreement on individual domain decisions $Prec_{dom}$ (1848 overall) was computed, ca. 86.85%.

A comparison between domain annotations of SRHs and ontological concepts (see Figure 4) suggests that annotating utterances with domains is an easier task for humans than annotating ontological concepts with the same information. A reason for this state of affairs might be that the high-level meaning of the utterance is defined through the context, whereas the meaning of an ontological concept is much less clear in the absence of context.

	$Prec_{mark}$		$Prec_{dom}$	
	Concepts	*SRHs*	*Concepts*	*SRHs*
Agreement	47.62%	92.5%	86.85%	98.92%

Table 4. Inter-annotator reliability for domain annotations.

2.3 Usage of the annotated data

In Section 2.2 we discussed the reasons for creating large corpora with annotated data. Such data is necessary to perform empirical research in the automatic language processing. At first, a research hypothesis is formulated. Then, it is translated into an annotation scheme and validated empirically

[4] Top-level concepts, e.g., *Type*, *Event* are typically not domain-specific. Therefore, they will not be assigned any domains.

by measuring the reliability of human annotations based on the corresponding annotation scheme. Additionaly, human performance indicates the *upper boundary*, also called *ceiling*, in evaluating computer programs.

The data from *Hidden Operator Tests* was used for the design and evaluation of some language processing algorithms. Following the common distinction in the evaluation practices, we differentiate between *intrinsic* and *extrinsic* evaluation. E.g., the semantic coherence scoring algorithm was evaluated intrinsically for scoring sets of ontological concepts in terms of their semantic coherence [6]. In this kind of evaluation, the output of the program is compared directly with the gold standard produced by the annotators. Extrinsic evaluation measures how a particular program contributes to the overall performance of the system. In our case, we employed the system in order to determine the best SRH in the output of the automatic speech recognizer [7].

A further application of the annotated data were a set of experiments directed at automatically assigning the domains to speech recognition hypotheses [16]. The algorithm for determining the domains of *conceptual representations* and their underlying speech recognition hypotheses employs a knowledge source called *Domain Model*. Two kinds of *Domain Models* were produced on the basis of the annotated data. The first *Domain Model* was derived by means of statistical analysis of speech recognition hypotheses in *Dataset 1* annotated for their domains. The second *Domain Model* was established through the direct annotation of concepts with respect to domains in the corresponding annotation experiments. Finally, the quality of both *Domain Models* and the domain recognition algorithm was evaluated intrinsically on the basis of annotated *Dataset 2*.

3 Collection and Usage of the *Wizard and Operator Test* Data

The main goal of the second data collection and its experimental setup was to enable precise analyses of the differences in the communicative behaviors of the various interlocutors, i. e., human-human, human-computer and computer-human interaction. The setup of the experiment was, therefore, designed to enable the control of various factors. Most important factors were the technical performance (e. g., latency times), the pragmatic performance (e. g., understanding vs. non-understanding of the user utterances) and the communicative behavior of the simulated system. They were to be adjustable to resemble the state of the art dialogue systems, such as SmartKom. These factors can, of course, also be adjusted to simulate potential future capabilities of dialogue systems and test their effects.

3.1 Data collection setup

For conducting the experiments a new paradigm for collecting telephone-based dialogue data, called *Wizard and Operator Test* (WOT) - described by [11] - was employed. This procedure represents a simplification of classical end-to-end experiments, as it is - much like *Wizard-of-Oz* (WoZ) experiments - conductable without the technically very complex use of a real conversational system. As post-experimental interviews showed, this did not limit the feeling of *authenticity* regarding the simulated conversational system by the human subjects. The WOT setup consists of two major phases that begin after subjects have been given a set of tasks to be solved with the telephone-based dialogue system:

- in `Phase 1` the human assistant is acting as a wizard who is simulating the dialogue system, much like in WoZ experiments, by operating a speech synthesis interface,
- in `Phase 2`, which starts immediately after a system breakdown has been simulated by means of beeping noises transmitted via the telephone, the human assistant is acting as a **human** operator asking the subject to continue with the tasks.

During the experiment the subject and the assistant were in separate rooms. Communication between both was conducted via telephone, i. e., for the user only a telephone was visible next to a radio microphone for the recording of the subject's linguistic expressions. As shown in Figure 5, the assistant/operator room featured a telephone as well as two computers - one for the speech synthesis interface and one for collecting all audio streams; loudspeakers were also present for feeding the speech synthesis output into the telephone and a microphone for the recording of the synthesis and operator output. With the help of an audio mixer all linguistic data were recorded time synchronously and stored in one audio file. The assistant/operator acting as the computer system communicated by selecting fitting answers for the subject's request from a prefabricated list which were returned via speech synthesis through the telephone. Beyond that it was possible for the assistant/operator to communicate over telephone directly with the subjects when acting as the human operator.

The experiments were conducted with an English setup, subjects and assistants at the International Computer Science Institute in Berkeley, USA, and with a German setup, subjects and assistants at the European Media Laboratory in Heidelberg, Germany. Both experiments were otherwise identical, with 22 sessions recorded in each. At the beginning of the WOT, the test manager told the subjects that they were testing a novel telephone-based dialogue system that supplies touristic information on the city of Heidelberg. In order to avoid the usual paraphrases of tasks worded too specifically, the manager gave the subjects an overall list of 20 very general touristic activities, such as *visit museum* or *Museum besuchen*, from which each subject had to

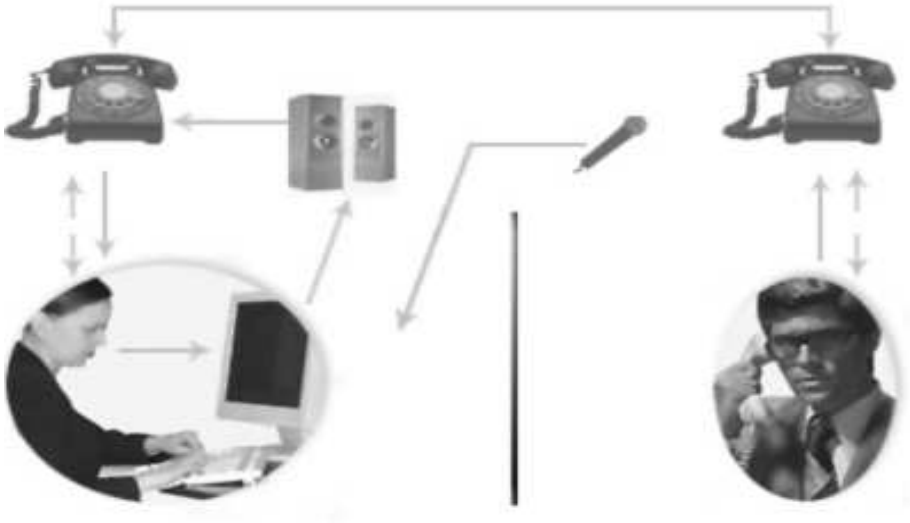

Fig. 5. Communication in `Phase 1` goes from synthesized speech out of the loudspeakers into the operator room (left side) telephone to the subject room (right side) and in `Phase 2` directly via the telephone between operator and subject.

pick six tasks which had to be solved in the experiment. The manager then removed the original list, dialed the system's number on the phone and exited from the room after handing over the telephone receiver. The subject was always greeted by the system's standard opening ply: *Welcome to the Heidelberger tourist information system. How can I help you?* After three tasks were finished - whether successful or not - the assistant simulated the system's break down and entered the line by saying *Excuse me, something seems to have happened with our system, may I assist you from here on* and finishing the remaining three tasks with the subjects.

Table 5 gives an overview of the data collected in the WOT experiments.

	Dataset E	*Dataset G*
Number of dialogues	22	22
Dialogue average length (min.)	5:53	4:57
Human-Operator average length (min.)	2:30	1:52
Human-Computer average length (min.)	3:23	2:59

Table 5. Descriptive corpus statistics of the English data (Dataset E) and the German data (Dataset G)

Overall the subjects featured approximately proportional mixtures of gender (25m, 18f), age (12< >71) and computer expertise. The average dialogue (German and English) consisted of 12.3 turns (composed of an average of 8.3 turn per dialogue in `Phase 1` and 16,3 in `Phase 2`).

3.2 Measurements on the data

This data enables analyses of the datasets that are language-specific as well as cross-linguistic analyses on the entire data, such as presented by [11]. In their analyses pauses (i.e. silences over 1 second) not caused by system latency times, overlaps in speech, dialogue structuring signals (e.g., *well*, *yes* or *OK*) and feedback-channeling signals [19] were in part manually and in part automaticall tagged and measured, yielding the results shown in Table 6.

	Dataset E	*Dataset G*
Dialogue average turns	14,33	10,28
Human-Operator average turns	21.25	11.35
Human-Computer average turns	7.4	9.2
Pauses	115	89
Pauses Human-Operator	21	10
Pauses Human-Computer	94	79
Overlaps	92	56
Overlaps Human-Operator	88	49
Overlaps Human-Computer	4	7
Dialogue-structuring Signals	292	317
Signals Human-Operator	202	225
Signals Human-Computer	90	112
Feedback Particles	43	153
Feedback Human-Operator	43	135
Feedback Human-Computer	0	18

Table 6. Measurements performed on the English data (Dataset E) and on the German data (Dataset G)

As the primary effects of the human-directed language exhibited by today's conversational dialogue systems, the data collected in these experiments clearly show that:

- dialogue efficiency decreases significantly even beyond the effects caused by latency times,
- the human interlocutor ceases in the production of feedback signals, but still attempts to use his or her turn signals for marking turn boundaries - which, however, remain ignored by the system,
- the increases in the amount of pauses is caused by waiting and uncertainty effects, which are also manifested by missing overlaps at turn boundaries.

4 Collection and Usage of the *Field Operative Test* Data

In an initial data collection we asked American native speakers to imagine that they are tourists in Heidelberg, Germany, equipped with a small, personal computer device that understands them and can answer their questions [12]. Among tasks from hotel and restaurant domains subjects also had to ask for directions to specific places. In the corpus we find 128 instances of instructional requests out of a total of roughly 500 requests from 49 subjects. The types and occurrences of these categories are in Table 7.

Type	*Example*	Occurrences
(A) How interrogatives	*How do I get to the Fischergasse*	38
(B) Where interrogatives	*Where is the Fischergasse*	37
(C) What/which interrogatives	*What is the best way to the castle*	18
(D) Imperatives	*Give me directions to the castle*	12
(E) Declaratives	*I want to go to the castle*	12
(F) Existential interrogatives	*Are there any toilets here*	8
(G) Others	*I do not see any bus stops*	3

Table 7. Request types and occurrences

As can be seen by looking at category (B), *Where interrogatives* are pragmatically ambiguous, i.e. they request either spatial localizations or spatial instructions. In order to perfrom an empirical study examining this phenomena under realistic conditions we perfromed the *Field-Operative Test* described below.

4.1 Data collection setup

Based on the results from these observations we conducted a *Field Operative Test* (FOT) in which field operatives asked people on the street using interrogatives of type A, B, C, D, E and F. E.g., `Where interrogatives` such as *Excuse me, can you tell me where the cinema Europa is* or `How interrogative` such as *How do I get to the castle* or `Existential Interrogatives` such as *Is there an ATM here.* The passerby's responses were not recorded as that would have required their permission and thwarted a *natural* response. We logged and varied several factors shown in Table 8.

We asked 366 subjects and their responses were, then, immediately categorized and logged according to their type, i.e. spatial descriptions, instructions, questions, etc. and their specific features, e.g. suggested means of transport (by bus, foot etc).

4.2 Usage of the field data

This set of collected and categorized data was used to train classifiers and extract decision trees via standard machine learning algorithms. For example,

Factors	Values
the goal object	the castle, city hall, school, discotheque, cinema, bank (ATM), clothing store
the time of day	morning, afternoon, evening
the proximity to the goal object	near , medium, far
the approximate age group	young, middle, old
the gender of the subjects	male, female
the weather conditions	warm, cold, rainy, dry
the operative's means of transportation	on foot, bicycle
the baggage situation of the question	with, without luggage
the accessibility of the goal object	open, closed

Table 8. Request types and occurrences

as reported by [13], the results of generating and rules applying a c4.5 learning algorithm (cf. [17]), show that:

- if the object is currently closed, e.g. a discotheque or cinema in the morning, almost 90% of the *Where interrogatives* are answered by means of localizations, a few subjects asked whether we actually wanted to go there now, and one subject gave instructions.
- if the object is currently open, e.g. a store or ATM machine in the morning, people responded with instructions, unless - and this we did not expect - the goal object is near and can be localized by means of a reference object that is within line of sight.

5 Conclusions

We described three experimental paradigms for the collection of data employed in the dialogue system design. We showed that issues in dialogue processing can be formulated as empirical problems, translated to annotation schemata and annotated corpora. The corpora can be employed as training data for statistical models and as a basis for symbolic models, such as domain models. A further advantage of our approach is a straightforward possibility for thorough evaluations of the developed algorithms. This is especially important for improving the quality and reliability of intuitive dialogue systems.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research and Technology (BMBF) as part of the SmartKom project under Grant

01 IL 905C/0 and by the Klaus Tschira Foundation. The authors would like to thank Dr. Michael Strube, Nicola Kaiser, Stefani Nellen, Florian Hillenkamp, Klaus Rüggenmann and Christof Müller for their help in designing and conducting the experiments.

References

1. James F. Allen, Bradford Miller, Eric Ringger, and Teresa Sikorski. A robust system for natural spoken dialogue. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 62–70, Santa Cruz, USA, 1996.
2. Gerard Bailly, Nick Campbell, and Bernd Möbius. ISCA Special Session: Hot topics in speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 37–40, Geneva, Switzerland, 2003.
3. Jean Carletta. Assessing agreement on classification tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 1996.
4. R.V. Cox, C.A. Kamm, L.R. Rabiner, J. Schroeter, and J.G. Wilpon. Speech and language processing for next-millenium communications services. *Proceedings of the IEEE*, 88(8):1314–1337, 2000.
5. Charles Darves and Shannon Oviatt. Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, U.S.A., 2002.
6. Iryna Gurevych, Rainer Malaka, Robert Porzel, and Hans-Peter Zorn. Semantic coherence scoring using an ontology. In *Proceedings of the HLT-NAACL 2003 Conference*, pages 88–95, Edmonton, CN, 2003.
7. Iryna Gurevych and Robert Porzel. Using knowledge-based scores for identifying best speech recognition hypotheses. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 77 – 81, Chateau-d'Oex-Vaud, Switzerland, 28-31 August 2003.
8. Iryna Gurevych, Robert Porzel, and Michael Strube. Annotating the semantic consistency of speech recognition hypotheses. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 46–49, Philadelphia, USA, July 2002.
9. Christoph Müller and Michael Strube. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, Japan, 4-5 July 2003.
10. Christoph Müller and Michael Strube. A tool for multi-level annotation of language data. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck)*, pages 199–200, Wallerfangen, Germany, 4-6 September 2003.
11. Robert Porzel and Manja Baudis. The Tao of HCI: Towards felicitous human-computer interaction. In *Proceedings of HLT-NAACL Conference*, Boston, USA, 2004. To appear.
12. Robert Porzel and Iryna Gurevych. Towards context-adaptive utterance interpretation. In *Proceedings of the 3nd SIGdial Workshop on Discourse and Dialogue,* Philadelphia, USA, July 2002, pages 90–95, 2002.

13. Robert Porzel and Iryna Gurevych. Contextual coherence in natural language processing. In P. Blackburn, C. Ghidini, R. Turner, and F. Giunchiglia, editors, *Modeling and Using Context*. LNAI 2680, Springer, Berlin, 2003.
14. Stefan Rapp and Michael Strube. An iterative data collection approach for multimodal dialogue systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 661–665, Las Palmas, Spain, 2002.
15. Laurent Romary, Michael Strube, and David Traum. Best practice in empirically-based dialogue research. 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck), http://www.coli.uni-sb.de/diabruck/pages/tutorial.htm, 4-6 September 2003.
16. Klaus Rüggenmann and Iryna Gurevych. Assigning domains to speech recognition hypotheses. Submitted, 2004.
17. Patrick Henry Winston. *Artificial Intelligence*. Addison-Wesley, 1992.
18. Robin Wooffitt, Nigel Gilbert, Norman Fraser, and Scott McGlashan. *Humans, Computers and Wizards: Conversation Analysis and Human (Simulated) Computer Interaction*. Brunner-Routledge, London, 1997.
19. V Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578, Chicago, Illinois, April 1970.
20. Magdalena Zoeppritz. Computer talk? Technical report, IBM Scientific Center Heidelberg Technical Report 85.05, 1985.