

# A Spoken Language Interpretation Component for a Robot Dialogue System

Enes Makalic, Ingrid Zukerman, Michael Niemann

Faculty of Information Technology, Monash University, Clayton, Australia  
{Enes.Makalic, Ingrid.Zukerman, Michael.Niemann}@infotech.monash.edu.au

## Abstract

The *DORIS* project aims to develop a spoken dialogue module for an autonomous robotic agent. This paper examines the techniques used by *Scusi?*, the speech interpretation component of *DORIS*, to postulate and assess hypotheses regarding the meaning of a spoken utterance. The results of our evaluation are encouraging, yielding good interpretation performance for utterances of different types and lengths.

**Index Terms:** speech understanding, probabilistic dialogue system, home environment

## 1. Introduction

The *DORIS* project aims to develop a spoken dialogue module for an autonomous robotic agent which supports the generation of responses that require physical as well as dialogue actions. In this paper, we describe *Scusi?*, *DORIS*'s language interpretation component, focusing on the techniques used to postulate and assess hypotheses regarding the meaning of a spoken utterance.

A language interpretation component should be able to postulate promising interpretations, and decide whether there is a clear winner or several likely candidates to be passed to the dialogue system. In order to support these capabilities, a spoken language interpretation system should (1) maintain multiple interpretations, and (2) apply a ranking process to assess the relative merit of each interpretation. *Scusi?* does this, employing (1) a multi-stage interpretation mechanism, where each stage maintains multiple options; and (2) a probabilistic mechanism which ranks interpretations according to their probability of matching the speaker's intention (§ 2). The system is highly modular, and integrates a probabilistic attribute comparison module for the disambiguation of referring expressions.

This paper is organized as follows: § 2 outlines the interpretation process and discusses the estimation of the probability of an interpretation. § 3 details our evaluation. Related research appears in § 4, and concluding remarks in § 5.

## 2. Interpretation Process

*Scusi?* processes spoken input in three stages: speech recognition, parsing and semantic interpretation (Fig. 1(a)). In the first stage, it runs Automatic Speech Recognition (ASR) software (Microsoft Speech SDK 5.1) to generate candidate texts from a speech signal. Each text is assigned a score that reflects the probability of the words given the speech wave. The second stage applies Charniak's probabilistic parser (<http://ftp.cs.brown.edu/pub/nlparser/>) to generate parse trees from the texts. The parser generates up to  $N$  ( $= 50$ ) parse trees for each text, associating each parse tree with a probability.

During semantic interpretation, parse trees are successively mapped into two representations based on Concept Graphs [1]: first *Uninstantiated Concept Graphs (UCGs)*, and then *Instantiated Concept Graphs (ICGs)*. UCGs are obtained from parse

trees deterministically – one parse tree generates one UCG. A UCG represents syntactic information, where the concepts correspond to the words in the parent parse tree, and the relations are derived from syntactic information in the parse tree and prepositions. Each UCG can generate many ICGs. This is done by nominating different instantiated concepts and relations from the system's knowledge base as potential realizations for each concept and relation in a UCG.

Fig. 1(b) illustrates the generation of an ICG for the request "leave the blue mug on the table in the corner". The noun 'mug' in the parse tree is mapped to the concept *mug* in the UCG, and then to the instantiated concept *mug03* in the ICG. The preposition 'on' in the parse tree is mapped to the relation *on* in the UCG, and then to the relation *Destination* in the ICG. Noun modifiers, such as colour and size, are treated as features to be matched to those of objects in the domain (§ 2.2).

Our interpretation algorithm applies a selection-expansion cycle to build a search graph, where each level of the graph corresponds to one of the above stages of the interpretation process (Fig. 1(a)). In each selection-expansion cycle, our algorithm selects an option for consideration (speech wave, textual ASR output, parse tree or UCG), and expands this option to the next level of interpretation. When an option is expanded, a single highly ranked candidate is returned for this next level (later expansions return lower ranked candidates in turn). For example, when we expand a text, the parser returns the next most probable parse tree for this text.

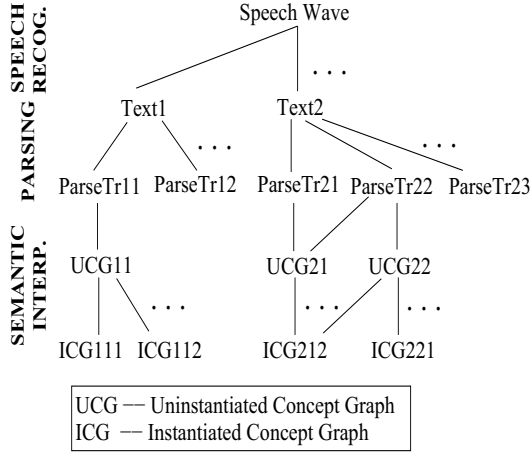
The consideration of all possible options at each stage of the interpretation process is computationally intractable. *Scusi?* uses two computational devices to generate interpretations in real time: (1) an anytime algorithm, and (2) a processing threshold. The anytime algorithm ensures that the system can return a list of ranked interpretations at any point after completing an expansion. The thresholding approach is used to prevent the consideration of unpromising alternatives. When the probability of the next child of a parent node  $n$  drops below a threshold  $Thr$  relative to the probability of the most probable child of  $n$ , no additional expansions of  $n$  are considered.

### 2.1. Estimating the probability of an ICG

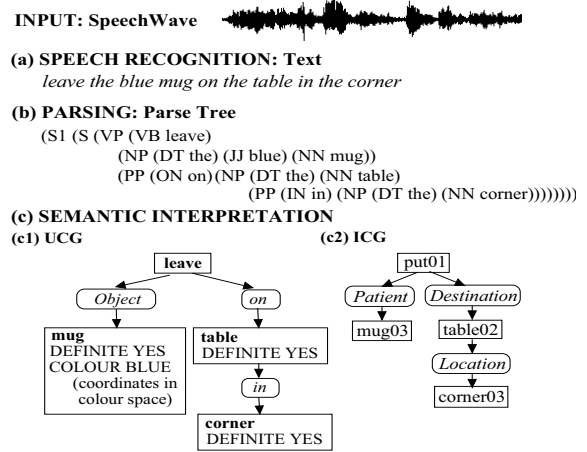
*Scusi?* ranks candidate ICGs according to their probability of being the intended meaning of a spoken utterance. Given a speech signal  $W$  and a context  $C$ , the probability of an ICG  $I$  is represented as follows.

$$\Pr(I|W, C) \propto \sum_{\Lambda} \Pr(I|U, C) \cdot \Pr(U|P) \cdot \Pr(P|T) \cdot \Pr(T|W) \quad (1)$$

where  $U$ ,  $P$  and  $T$  denote a UCG, parse tree and text respectively. The summation is taken over all possible paths  $\Lambda = \{P, U\}$  from a parse tree to the ICG, because a UCG and an ICG can have more than one parent (Fig. 1(a)). The ASR and the parser return an estimate for  $\Pr(T|W)$  and  $\Pr(P|T)$  re-



(a) Stages of the interpretation process



(b) Sample structures for the interpretation stages

Figure 1: *Scusi?*'s spoken language interpretation process

spectively; and  $\Pr(U|P) = 1$ , since the process of generating a UCG from a parse tree is deterministic.

The estimation of  $\Pr(I|U, C)$  is described in detail in [2]. Here we present the final equation obtained for  $\Pr(I|U, C)$ , and outline the ideas involved in its calculation.

$$\Pr(I|U, C) \approx \prod_{k \in I} \Pr(u|k) \Pr(k|k_p, k_{gp}) \Pr(k|C) \quad (2)$$

where  $k$  is an instantiated node in ICG  $I$ ,  $u$  is the corresponding node in UCG  $U$ ,  $k_p$  is the parent node of  $k$ , and  $k_{gp}$  the grandparent node of  $k$ .

- $\Pr(u|k)$  is the “match probability” between the features of node  $k$  in ICG  $I$ , and the specifications for the corresponding node  $u$  in UCG  $U$ , e.g., how similar an object in the room is to the “blue mug” (§ 2.2).
- $\Pr(k|k_p, k_{gp})$  is the structural probability of ICG  $I$ , where structural information is simplified to node trigrams in the ICG (e.g., with reference to Fig. 1(b), whether the *Location* of *table02* is indeed *corner03*).
- $\Pr(k|C)$  is the probability of a concept in light of the context, which at present includes only domain knowledge.

## 2.2. Probabilistic Feature Comparison

This section describes how *Scusi?* estimates the “match probability”,  $\Pr(u|k)$ , between intrinsic features of a UCG node  $u$  and an ICG node  $k$ . *Scusi?* currently handles three intrinsic features: lexical item, colour and size. For instance, “the big red cup” specifies the features `lexical_item="cup"`, `colour="red"`, and `size="big"`. These features are compared to the features of objects in the domain in order to propose suitable candidates for “the big red cup” when an ICG is created from a parent UCG (semantic interpretation stage). In agreement with [3], lexical item and colour are considered *absolute* features, and size a *relative* feature (its value depends on the size of other candidates).

At present, we make the following simplifying assumptions: (1) the robot is co-present with the user and the possible referents of an utterance; and (2) the robot has an unobstructed view of the objects in the room and up-to-date information about these objects (this information could be obtained through a scene analysis system activated upon entering a room). These assumptions obviate the need for planning physical actions, such as moving to get a better view of some objects.

### 2.2.1. Building a list of candidate instantiated concepts.

For each node  $u$  in a UCG  $U$ , the task is to construct a list of candidate instantiated concepts  $k \in \mathcal{K}$  that are a reasonable match for  $u$  ( $\mathcal{K}$  is the knowledge base of objects in the domain). This list is built as follows.

1. Estimate  $\Pr(u|k)$ , the probability of the match between the features of  $u$  and those of an instantiated concept  $k$ .
2. Rank the candidates in descending order of probability.
3. Filter out the candidates whose probability falls below a threshold (first lexical threshold, next colour, then size).

### 2.2.2. Estimating the probability of a match.

The probability of the match between a node  $u$  in UCG  $U$  and a candidate instantiated concept  $k \in \mathcal{K}$  is estimated as follows.

$$\Pr(u|k) = \Pr(\mathbf{u}_{f_1}, \dots, \mathbf{u}_{f_p} | \mathbf{k}_{f_1}, \dots, \mathbf{k}_{f_p}) \quad (3)$$

where  $(f_1, \dots, f_p) \in \mathcal{F}$  are the features specified with respect to node  $u$ ,  $\mathcal{F}$  is the set of features allowed in the system,  $\mathbf{u}_{f_i}$  is the value of the  $i$ -th feature of UCG node  $u$ , and  $\mathbf{k}_{f_i}$  is the value of this feature for the instantiated concept  $k$ .

Assuming that the individual features of a node are independent, the probability that an instantiated concept  $k$  matches the specifications in a UCG node  $u$  can be rewritten as

$$\Pr(u|k) = \prod_{i=1}^p \Pr(\mathbf{u}_{f_i} | \mathbf{k}_{f_i}) \quad (4)$$

We use a linear distance function  $h: \mathbb{R}^+ \rightarrow [0, 1]$  to map the outcome of a match of feature  $f$  to the probability space. Specifically,

$$\Pr(\mathbf{u}_f | \mathbf{k}_f) = h_f(\mathbf{u}_f, \mathbf{k}_f) \quad (5)$$

The calculation of Equation 5 for the intrinsic features supported by *Scusi?* is presented in the following sections.

### 2.2.3. Lexical item

We employ the Leacock and Chodorow [4] similarity measure, denoted  $LC$ , to compute the similarity between the lexical feature of  $u$  and  $k$  (this measure yielded the best results among those we investigated). The  $LC$  similarity score, denoted  $s_{LC}$ , is converted to a probability by applying the following function.

$$\Pr(\mathbf{u}_{\text{lex}} | \mathbf{k}_{\text{lex}}) = h_{\text{lex}}(s_{LC}(\mathbf{u}_{\text{lex}}, \mathbf{k}_{\text{lex}})) = \frac{s_{LC}(\mathbf{u}_{\text{lex}}, \mathbf{k}_{\text{lex}})}{s_{\text{max}}}$$

where  $s_{\text{max}}$  is the highest possible  $LC$  score.

### 2.2.4. Colour

The colour model chosen for *Scusi?* is the CIE 1976  $(L, a, b)$  colour space, which has been experimentally shown to be approximately perceptually uniform [5]. The  $L$  coordinate represents brightness ( $L = 0$  denotes black, and  $L = 100$  white),  $a$  represents position between green ( $a < 0$ ) and red ( $a > 0$ ), and  $b$  position between blue ( $b < 0$ ) and yellow ( $b > 0$ ). The range of  $L$  is  $[0, 100]$ , while for practical purposes, the range of  $a$  and  $b$  is  $[-200, 200]$ . Thus, the probability of a colour match between a UCG concept  $u$  and an instantiated concept  $k$  is

$$\Pr(\mathbf{u}_{\text{color}}|\mathbf{k}_{\text{color}}) = h_{\text{color}}(\mathbf{u}_{\text{color}}, \mathbf{k}_{\text{color}}) = 1 - \frac{ED(\mathbf{u}_{\text{color}}, \mathbf{k}_{\text{color}})}{d_{\text{max}}}$$

where  $ED$  is the Euclidean distance between the  $(L, a, b)$  coordinates of the colour specified for  $u$  and the  $(L, a, b)$  coordinates of the colour of  $k$ , and  $d_{\text{max}}$  is the maximum Euclidean distance between two colours (=574.5).

### 2.2.5. Size

Unlike lexical item and colour, size is considered a relative feature, i.e., the probability of a size match between an object  $k \in \mathcal{K}$  and a UCG concept  $u$  depends on the sizes of all suitable candidate objects in  $\mathcal{K}$  (those that exceed the thresholds for lexical and colour comparisons). The highest probability for a size match is then assigned to the object that best matches the required size, while the lowest probability is assigned to the object which has the worst match with this size.

This requirement is achieved by the following function, which like Kelleher’s [6] pixel-based mapping, performs a linear mapping between  $\mathbf{u}_{\text{size}}$  and  $\mathbf{k}_{\text{size}}$ .

$$\Pr(\mathbf{u}_{\text{size}}|\mathbf{k}_{\text{size}}) = h_{\text{size}}(\mathbf{u}_{\text{size}}, \mathbf{k}_{\text{size}}) = \begin{cases} \frac{\alpha k_{\text{size}}}{\max_i \{k_{\text{size}}^i\}} & \text{if } \mathbf{u}_{\text{size}} \in \{\text{‘large’/‘big’/...}\} \\ \frac{\alpha \min_i \{k_{\text{size}}^i\}}{k_{\text{size}}} & \text{if } \mathbf{u}_{\text{size}} \in \{\text{‘small’/‘little’/...}\} \end{cases}$$

where  $\alpha$  is a normalizing constant, and  $k_{\text{size}}^i$  is the size of candidate object  $k^i$  (e.g.,  $k_{\text{size}}^{\text{mug}03} = 0.9\text{dm}^3$ ). This formula is adapted for individual dimensions, e.g., length.

### 2.2.6. Combining Feature Scores

To determine how features are used in our domain, we conducted a survey where people were asked to refer to household objects laid out in a space. Our survey found that people often present features that are not strictly necessary to identify an item, and use features in the following order of frequency: *type*  $\succ$  *absolute adjectives*  $\succ$  *relative adjectives*, where colour is an absolute concept and size is a relative concept. This finding agrees with the feature ranking observed by Dale and Reiter [3].

This prompted us to incorporate a weighting scheme into Equation 4, where the features are weighted according to their usage in referring expressions. That is, higher ranking or more frequently used features are assigned a higher weight than lower ranking or less frequently used features. Specifically, given a match probability  $\Pr(\mathbf{u}_{f_i}|\mathbf{k}_{f_i})$  and a weight  $w_{f_i}$  for feature  $f_i$ , the adjusted match probability is

$$\Pr'(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) = \Pr(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) \times w_{f_i} + \frac{1}{2}(1 - w_{f_i}) \quad (6)$$

where  $0 < w_{f_i} \leq 1$ . The effect of this mapping is that features with high weights have a wide range of probabilities (and hence a substantial influence on the match probability of an object), while features with low weights have a narrow probability range (and a reduced influence on match probability). For example, if  $w_{f_i} = 0.6$ ,  $0.2 \leq \Pr'(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) \leq 0.8$ , while if  $w_{f_i} = 0.8$ ,  $0.1 \leq \Pr'(\mathbf{u}_{f_i}|\mathbf{k}_{f_i}) \leq 0.9$ .

	# Gold refs with prob in top 1	top 3	Average adj rank (rank)	Avg # to Gold refs (iters)
BASELINE	26	47	1.33 (1.25)	2.8 (22)
DR	33	54	0.70 (0.63)	2.7 (20)
FREQ	31	51	0.81 (0.75)	2.8 (21)
<b>Total</b>	<b>56</b>	<b>56</b>		

Table 1: *Scusi?*’s interpretation of referring expressions

## 3. Evaluation

We conducted two experiments to evaluate our system. In the first experiment, we determined optimal feature weights, and in the second experiment with assessed *Scusi?*’s overall interpretation performance. In both experiments, *Scusi?* was set to generate at most 300 interpretations in total (including texts, parse trees, UCGs and ICGs) for each utterance in the test set. An interpretation was deemed correct if it matched the speaker’s intention, which in turn was represented by one or more Gold ICGs. Multiple Gold ICGs were allowed if several objects in the domain matched a specified object, e.g., “get a mug”.

In the first experiment, we constructed six “worlds”, each comprising 13–26 objects. We then composed a set of 7–15 descriptions for each world (e.g., “the long yellow tray”), yielding a total of 56 referring expressions. The objects, and their size and colour, were chosen so that they had similar features, e.g., a stool may also be called ‘seat’ or ‘chair’, and there were objects in different shades of the same colour. To establish which objects should be considered the Gold standard, pictures of the six worlds and their lists of descriptions were shown to two human taggers. The taggers independently identified one or more objects which best corresponded to each description (inter-tagger agreement was  $\kappa = 0.86$ ). When the taggers disagreed, Gold standards were derived through consensus-based annotation [7].

We considered the following schemes for assigning weights to  $(w_{\text{lex}}, w_{\text{color}}, w_{\text{size}})$  in Equation 6.

- BASELINE (1, 1, 1).
- DR (1, 0.8, 0.6) – based on the feature ordering in [3].
- FREQ (1, 0.79, 0.39) – based on the frequencies observed in our formative survey (§ 2.2.6).

Table 1 summarizes our results. Column 1 shows the weighting scheme. Columns 2 and 3 show how many of the descriptions had Gold referents whose probability was the highest (top 1) or among the three highest (top 3). The average *adjusted rank* and *rank* of the Gold referent appear in Column 4. The rank of a referent  $r$  is its position in a list sorted in descending order of probability (starting from position 0), such that all equiprobable referents are deemed to have the same position. The adjusted rank of a referent  $r$  is the mean of the actual positions of all referents that have the same probability as  $r$ . Column 5 indicates the average number of referents created and iterations performed until the Gold referent was found.

The performance of the baseline differs from that of the weighted schemes ( $p < 0.03$ ), but the difference between the weighted schemes is not statistically significant.<sup>1</sup> Since reference disambiguation accuracy is improved by assigning weights to the probabilities of intrinsic features, we employed the DR weights in our next experiment (they seem to perform marginally better than our frequency-based weights).

In the second experiment, we assessed the overall performance of our approach, and determined the impact of various thresholds on interpretation performance. As a baseline, we

<sup>1</sup>Sample paired t-tests were used for all statistical tests.

	# Gold ICGs with prob in top 1	# Gold ICGs with prob in top 3	Average adj rank(rank)	Not found	Avg # to Gold ICGs (iters)
<b>BASELINE</b>	53	53	0 (0)	47	0 (4)
No Thrsh	69	82	3.85 (1.15)	7	9 (38)
10%	67	81	2.63 (0.91)	8	8 (37)
20%	70	83	2.47 (0.87)	7	8 (39)
50%	70	84	2.37 (0.81)	7	8 (37)
80%	70	85	2.31 (0.80)	7	8 (37)
90%	70	85	2.31 (0.78)	7	8 (37)
<b>Total</b>	100	100			

Table 2: *Scusi?*'s overall interpretation performance

executed a beam search where only the best result from each interpretation stage was considered. The evaluation test set comprised 100 utterances: 43 declarative (e.g., “the book is on the desk”, “in the kitchen”, “the red mug”) and 57 imperative (e.g., “open the door”). These utterances were based on interactions between users and a “robot” (enacted by one of the authors) in a virtual home scenario. The utterances were chosen to test *Scusi?*'s ability to identify target objects (the intended book, mug, table, etc), and its ability to handle phenomena such as synonyms (e.g., “wash” and “clean”) and homonyms (e.g., “leave the mug on the table” versus “leave the room”). Average utterance length was 8.5 words, with a maximum length of 12 words. Gold ICGs were manually constructed by the speaker (one of the authors) from *Scusi?*'s knowledge base, which comprises 135 items (24 relations and 111 concepts).

Table 2 summarizes our results. The columns are similar to those in Table 1, except the “Not found” column, which indicate how many Gold ICGs were never generated. The results in Table 2 demonstrate that our approach outperforms the baseline approach ( $p < 0.05$ ). The number of top-ranked Gold ICGs and not found ICGs, and number of iterations to Gold are largely threshold invariant. However, the average rank of the Gold ICGs decreases (improves) as the threshold increases, which is consistent with the slight improvement in the number of top-3 ICGs. We also performed additional experiments that examined the effect of using a different threshold for each level of interpretation. However, the new scheme did not yield any improvement over a single system-wide threshold.

#### 4. Related Work

Many researchers have investigated numerical approaches to the interpretation of spoken utterances in dialogue systems, e.g., [8, 9, 10]. Pflieger *et al.* [8] employ modality fusion to combine hypotheses from different analyzers (linguistic, visual and gesture), and apply a scoring mechanism to rank the resultant hypotheses. They disambiguate referring expressions by choosing the first object that satisfies a ‘differentiation criterion’, hence their system does not handle situations where more than one object satisfies this criterion. He and Young [9] and Gorniak and Roy [10] apply a probabilistic approach to spoken language interpretation. All of these systems employ semantic grammars, while *Scusi?* uses generic, syntactic tools, and incorporates semantic- and domain-related information only in the final stage of the interpretation process. Knight *et al.* [11] compare the performance of a dialogue system based on a semantic grammar to that of a system based on a statistical language model and a robust phrase-spotting grammar. The latter performs better for relatively unconstrained utterances by users unfamiliar with the system. The probabilistic approach and intended users of our system are in line with this finding.

Kelleher [6] proposes a reference resolution algorithm that accounts for four attributes: lexical type, colour, size and loca-

tion, where the score of an object is estimated by a weighted combination of the visual and linguistic salience scores of each attribute. However, Kelleher limits the probabilistic comparison of features to size and location, and uses binary comparisons for lexical item and colour.

#### 5. Conclusion

We have described *Scusi?*, a spoken language interpretation system that maintains multiple options at each stage of the interpretation process, and ranks interpretations based on estimates of their posterior probability. As part of *Scusi?*, we presented a probabilistic reference disambiguation mechanism which considers intrinsic features of domain objects.

Our empirical evaluation shows that (1) *Scusi?* performs well for declarative and imperative utterances of varying length, with the Gold ICG(s) receiving one of the top three probabilities for most test utterances; (2) maintaining multiple interpretations yields a better performance than the baseline approach, but the threshold has no significant effect on *Scusi?*'s performance; and (3) reference disambiguation accuracy is improved by assigning weights to the probabilities of intrinsic features.

#### 6. References

- [1] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.
- [2] I. Zukerman, E. Makalic, M. Niemann, and S. George, “A probabilistic approach to the interpretation of spoken utterances,” Faculty of Information Technology, Monash University, Clayton, Victoria, Tech. Rep. 226, 2008.
- [3] R. Dale and E. Reiter, “Computational interpretations of the Gricean maxims in the generation of referring expressions,” *Cognitive Sci.*, vol. 18, no. 2, pp. 233–263, 1995.
- [4] C. Leacock and M. Chodorow, “Combining local context and WordNet similarity for word sense identification,” in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. MIT Press, 1998, pp. 265–285.
- [5] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi, “Empirical evaluation of dissimilarity measures for color and texture,” in *the 7th IEEE Int. Conf. on Computer Vision*, vol. 2, Kerkyra, Greece, 1999, pp. 1165–1172.
- [6] J. Kelleher, “Attention driven reference resolution in multimodal contexts,” *Artificial Intelligence Review*, vol. 25, pp. 21–35, 2006.
- [7] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *ICSLP-2002*, Denver, Colorado, 2002, pp. 2037–2040.
- [8] N. Pflieger, R. Engel, and J. Alexandersson, “Robust multimodal discourse processing,” in *the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany, 2003, pp. 107–114.
- [9] Y. He and S. Young, “A data-driven spoken language understanding system,” in *ASRU’03*, St. Thomas, US Virgin Islands, 2003.
- [10] P. Gorniak and D. Roy, “Probabilistic grounding of situated speech using plan recognition and reference resolution,” in *ICMI’05*, Trento, Italy, 2005, pp. 138–143.
- [11] S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin, “Comparing grammar-based and robust approaches to speech understanding: A case study,” in *Europeespeech 2001*, Aalborg, Denmark, 2001.