Evaluation of the *Scusi?* Spoken Language Interpretation System – A Case Study

Thomas Kleinbauer, Ingrid Zukerman and **Su Nam Kim**
Faculty of Information Technology, Monash University
Clayton, Victoria 3800, Australia

Abstract

We present a performance evaluation framework for Spoken Language Understanding (SLU) modules, focusing on three elements: (1) characterization of spoken utterances, (2) experimental design, and (3) quantitative evaluation metrics. We then describe the application of our framework to *Scusi?*— our SLU system that focuses on referring expressions.

1 Introduction

We present a performance evaluation framework for Spoken Language Understanding (SLU) modules, and describe its application to the evaluation of *Scusi?* — an SLU system that focuses on the interpretation of descriptions of household objects (Zukerman et al., 2008). Our contributions pertain to (1) the characterization of spoken utterances, (2) experimental design, and (3) quantitative evaluation metrics for an N-best list.

Characterization of spoken utterances. According to (Jokinen and McTear, 2010), "in diagnostic-type evaluations, a representative test suite is used so as to produce a system's performance profile with respect to a taxonomy of possible inputs". In addition, one of the typical aims of an evaluation is to identify components that can be improved (Paek, 2001). These two factors in combination motivate a characterization of input utterances along two dimensions: *accuracy* and *knowledge* (Section 4).

- ***Accuracy*** indicates whether an utterance describes an intended object precisely and unambiguously. For instance, when intending a blue plate, "the blue plate" is an accurate description if there is only one such plate in the room, while "the *green* plate" is inaccurate.
- ***Knowledge*** indicates how much the SLU module knows about different factors of the interpretation process, e.g., vocabulary or geometric relations. For instance, "CPU" in "the *CPU* under the desk"∗[1] is *Out of Vocabulary* (*OOV*) for *Scusi?*, and the "of" in "the picture *of* a face"∗ is an unknown relation.

The frequency of different values for these dimensions influence the requirements from an SLU system, and the components that necessitate additional resources, e.g., vocabulary extension.

Experimental design. It is generally accepted that an SLU system should exhibit reasonable behaviour by human standards. At present, in experiments that evaluate an SLU system's performance, people speak to the system, and the accuracy of the system's interpretation is assessed. However, this mode of evaluation, which we call *Generative*, does not address whether a system's interpretations are plausible (even if they are wrong). Thus, in addition to a Generative experiment, we offer an *Interpretive* experiment. Both experiments are briefly described below. Their implementation in our SLU system is described in Section 5.

- In the ***Interpretive*** experiment, trial subjects and the SLU system are addressees, and are given utterances generated by a third party. The SLU system's confidence in its interpretations is then compared with the preferences of the participants.
- In the ***Generative*** experiment, trial subjects are speakers, generating free-form utterances, and the SLU module and expert annotators are addressees. Gold standard interpretations for these descriptions are produced by annotators on the basis of their understanding of what was said, e.g., an ambiguous utterance has more than one correct interpretation. The SLU system's performance is evaluated on the basis of the rank of the correct interpretations.

[1] Examples from our trials are marked with asterisks (∗).

International Joint Conference on Natural Language Processing, pages 225–233,
Nagoya, Japan, 14-18 October 2013.

These two experiments, in combination with our characterization of spoken utterances, enable the comparison of system and human interpretations under different conditions.

Quantitative evaluation metrics. Automatic Speech Recognizers (ASRs) and parsers often return N-best hypotheses to SLU modules, while many SLU systems return only one interpretation (DeVault et al., 2009; Jokinen and McTear, 2010; Black et al., 2011). However, maintaining N-best interpretations at the semantic and pragmatic level enables a Dialogue Manager (DM) to examine more than one interpretation, and discover features that guide appropriate responses and support error recovery. This ranking requirement, together with our experimental design, motivates the following metrics (Section 6).

- For ***Interpretive*** experiments, we propose correlation measures, such as Spearman rank or Pearson correlation coefficient, to compare participants' ratings of candidate interpretations with the scores given by an SLU system.
- For ***Generative*** experiments, we provide a broad view of an SLU system's performance by counting the utterances that it *CantRepresent*, and among the remaining utterances, counting those for which a correct interpretation was *NotFound*. We obtain a finer-grained view using fractional variants of the Information Retrieval (IR) metrics *Recall* (Salton and McGill, 1983) and *Normalized Discounted Cumulative Gain* (*NDCG*) (Järvelin and Kekäläinen, 2002), which handle equiprobable interpretations in an N-best list. We also compute $@K$ versions of these metrics to represent the relation between rank and performance.

In the next section, we discuss related work, and in Section 3, we outline our system *Scusi?*. In Section 4, we present our characterization of descriptions, followed by our experimental design and evaluation metrics. The results obtained by applying our framework to *Scusi?* are described in Section 7, followed by concluding remarks.

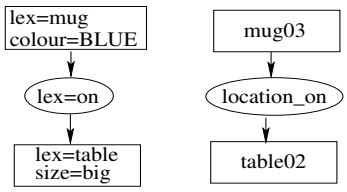
2 Related Work

As mentioned above, our contributions pertain to the characterization of spoken utterances, experimental design, and quantitative metrics.

Characterization of spoken utterances. Most evaluations of SLU systems characterize input utterances in terms of ASR *Word Error Rate* (*WER*), e.g., (Hirschman, 1998; Black et al., 2011). Möller (2008) provides a comprehensive collection of interaction parameters for evaluating telephone-based spoken dialogue services, which pertain to different aspects of an interaction, viz communication, cooperativity, task success, and spoken input. Our characterization of spoken utterances along the accuracy and knowledge dimensions is related to Möller's task success category. However, in our case, these features pertain to the context, rather than the task. In addition, our characterization is linked to system development effort, i.e., how much effort should be invested to address utterances with certain characteristics; and to evaluation metrics, in the sense that the assessment of an interpretation depends on the accuracy of an utterance, and takes into account the capabilities of an SLU system.

Experimental design. Evaluations performed to date are based on Generative experiments (Hirschman, 1998; Gandrabur et al., 2006; Thomson et al., 2008; DeVault et al., 2009; Black et al., 2011), which focus on correct or partially correct responses. They do not consider human interpretations for utterances with diverse characteristics, as done in our Interpretive trials.

Quantitative evaluation metrics. Most SLU system evaluations use IR-based metrics, such as recall, precision and accuracy, to compare the components of one interpretation of a perfect request to the components of a reference interpretation (Hirschman, 1998; Möller, 2008; DeVault et al., 2009; Jokinen and McTear, 2010). In contrast, we consider the rank of completely correct interpretations of perfect requests and partially correct interpretations of imperfect requests in an N-best list. Thomson *et al.* (2008) analyzed metrics for N-best lists, such as *Receiver Operator Characteristic*, *Weighted Semantic Error Rate* and *Normalized Cross Entropy* (Gandrabur et al., 2006); and offered the *Item Level Cross Entropy* (*ICE*) metric, which combines the confidence score and correctness of each of N-best interpretations. In this paper, we adapt IR-based metrics to handle equiprobable interpretations in an N-best list, and offer the *CantRepresent* and *NotFound* metrics to give a broad view of system performance. In the future, we intend to incorporate confidence/accuracy metrics, such ICE.



(a) UCG (b) ICG

Figure 1: Sample UCG and ICG for "the blue mug on the large table".

3 The *Scusi?* System

Scusi? is a system that implements an anytime, probabilistic mechanism for the interpretation of spoken utterances, focusing on a household context. It has four processing stages, where each stage produces multiple outputs for a given input, early processing stages may be probabilistically revisited, and only the most promising options at each stage are explored further.

The system takes as input a speech signal, and uses an ASR (Microsoft Speech SDK 6.1) to produce candidate texts. Each text is assigned a probability given the speech wave. The second stage applies Charniak's probabilistic parser (`bllip.cs.brown.edu/resources.shtml#software`) to syntactically analyze the texts in order of their probability, yielding at most 50 different parse trees per text. The third stage applies mapping rules to the parse trees to generate *Uninstantiated Concept Graphs* (*UCGs*) that represent the semantics of the utterance (Sowa, 1984). The final stage produces *Instantiated Concept Graphs* (*ICGs*) that match the concepts and relations in a UCG with objects and relations within the current context (e.g., a room), and estimates how well each instantiation matches its "parent" UCG and the context. For example, Figure 1(a) shows one of the UCGs returned for the description "the blue mug on the large table", and Figure 1(b) displays one of the ICGs generated for this UCG. Note that the concepts in the UCG have generic names, e.g., `mug`, while the ICG contains specific objects, e.g., `mug03` or `cup01`, which are offered as candidate matches for `lex=mug, color=blue`.

3.1 *Scusi?*'s capabilities

Scusi? aims to understand requests for actions involving physical objects (Zukerman et al., 2008). Focusing on object descriptions, *Scusi?* has a vocabulary of lexical items pertaining to objects, colours, sizes and positions. For object names, this vocabulary is expanded with synonyms and near synonyms obtained from WordNet (Fellbaum, 1998) and word similarity metrics from (Leacock and Chodorow, 1998). However, this vocabulary is not imposed on the ASR, as we do not want *Scusi?* to hear only what it wants to hear. In addition, *Scusi?* was designed to understand the colour and size of objects; the topological positional relations *on, in, near* and *at*, optionally combined with *center, corner, edge* and *end*, e.g., "the mug *near the center* of the table"; and the projective positional relations *in front of, behind, to the left/right, above* and *under* (topological and projective relations are discussed in detail in (Coventry and Garrod, 2004; Kelleher and Costello, 2008)). By "understanding a description" we mean mapping attributes and positions to values in the physical world. For instance, the CIE colour metric (CIE, 1995) is employed to understand colours, Gaussian functions are used to represent sizes of things compared to the size of an average exemplar, and spatial geometry is used to understand positional relations.

At present, *Scusi?* does not understand (1) *OOV* words, e.g., "the *opposite* wall"∗; (2) more than one meaning of polysemous positional relations, e.g., "*to the left of* the table"∗ as "*to the left and on* the table" as well as "*to the left and next to* the table"; (3) positional relations that are complex, e.g., "*in the left near corner of* the table"∗, or don't have a landmark, e.g., "the ball *in the center*"∗; and (4) descriptive prepositional phrases starting with "of" or "with", e.g., "the picture *of* the face"∗ and "the plant *with* the leaves"∗. However, contextual information sometimes enables the system to overcome OOV words. For example, *Scusi?* may return the correct ICG for "the *round* blue plate on the table" at a good rank.

Clearly, these problems can be solved by programming additional capabilities into our system. However, people will always say things that an SLU system cannot understand. Our evaluation framework can help distinguish between situations in which it is worth investing additional development effort, and situations for which other coping mechanisms should be developed, e.g., asking a clarification question or ignoring the unknown portions of an utterance (while being aware of the impact of this action on comprehension).

3.2 ASR capabilities

The WER of the ASR used by *Scusi?* is 30% when trained on an open vocabulary in combination with a small language model for our corpus.

This WER is consistent with the WER obtained in the 2010 Spoken Dialogue Challenge (Black et al., 2011). In addition to the obvious problem of misrecognized entities or actions, which yield OOV words, ASR errors often produce ungrammatical sentences that cannot be successfully parsed. For instance, one of the alternatives produced by the ASR for "the blue plate at the front of the table"∗ is "*to build played* at the front *door* the table". Further, disfluencies are often mis-heard by the ASR or cause it to return broken sentences.

4 Characterization of Spoken Utterances

When describing an object or action, speakers may employ a wrong lexical item, or use a wrong attribute. For instance, "the green *couch*"∗ was described when intending a green bookcase. In addition, when describing objects, speakers may under-specify them, e.g., ask for "the pink mug" when there are several such mugs; provide inconsistent specifications that do not match any object perfectly, yielding no candidates or several partial candidates, e.g., request "the large blue mug" when there is a large pink mug and a small blue mug; omit a landmark, e.g., "the ball in the center"∗; or employ words or constructs unknown to an SLU module, e.g., "the *exact* center"∗.[2] These situations, which affect the performance of an SLU system, are characterized along the following two dimensions: *accuracy* and *knowledge*.

- ***Accuracy*** **–** We distinguish between *Perfect* and *Imperfect* utterances. An utterance is perfect if it matches at least one object or action in the current context in every respect. In this case, an SLU module should produce one or more interpretations that match perfectly the utterance. If every object or action in the context mismatches an utterance at least in one aspect, the utterance is *imperfect*. In this case, we consider *reasonable* interpretations (that match the request well but not perfectly) to be the Gold standard. The ***number*** of Gold interpretations is an attribute of accuracy: an utterance may match (perfectly or imperfectly) 0, 1 or more than 1 interpretation.
- ***Knowledge*** **–** If all the words and syntactic constructs in an utterance are understood by an SLU module (Section 3.1), the utterance is deemed *known*, otherwise, it is *unknown*.

[2]People often over-specify their descriptions, e.g., "the large red mug" when there is only one red mug (Dale and Reiter, 1995). Such over-specifications are not problematic.

To illustrate these concepts, a description that contains only known words, and matches two objects in the context in every respect, is classified as *known-perfect*>1.

5 Experimental Design

We devised two experiments to assess an SLU system's performance: ***Interpretive***, where the participants and the SLU system are the addressees (Section 5.1), and ***Generative***, where the participants are the speakers and the SLU module is the addressee (Section 5.2).

In both experiments, we evaluate the performance of an SLU system on the basis of complete interpretations of an utterance, which in *Scusi?*'s case is a description. For example, given "the pink ball near the table", all the elements of an ICG must try to match this description and the context. That is, if `ball01` is pink, but it is *on* `table02`, the ICG `ball01`–*`location_near`*–`table02` will have a good description match but a bad reality match, while the opposite happens for ICG `ball01`–*`location_on`*–`table02`.

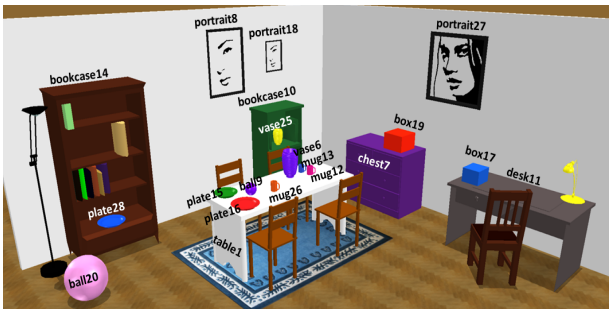
5.1 Interpretive trial

This experiment tests whether *Scusi?*'s understanding matches the understanding of a relatively large population under different accuracy conditions. We focus on imperfect and ambiguous descriptions, as they pose a greater challenge to people than perfect descriptions. The trial consists of a Web-based survey where participants were given a picture of a room and 9 descriptions generated by the authors (Figure 2). For each description, participants were asked to rate each of 20 labeled objects based on how well they match the description, where a rating of 10 denotes a "perfect match" and a rating of 0 denotes "no match".

Our Web survey was done by 47 participants, resulting in 47×20 scores for each description. These scores were averaged across participants, yielding a single score for each labeled object for each of our 9 descriptions.

5.2 Generative trial

In this experiment, trial subjects generated free-form, spoken descriptions to identify three designated objects in each of four scenarios. The scenarios, which were designed to test different functionalities of *Scusi?*, contain between 8 and 16 objects (Figure 3 shows two scenarios). The annotators provided the Gold standard interpretations for

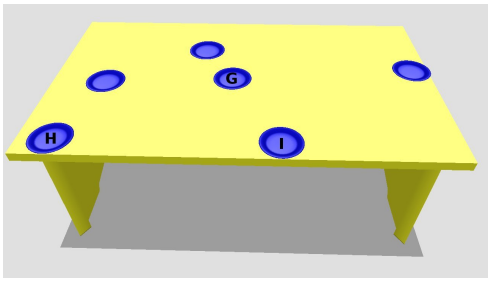


(a) Room with labeled objects

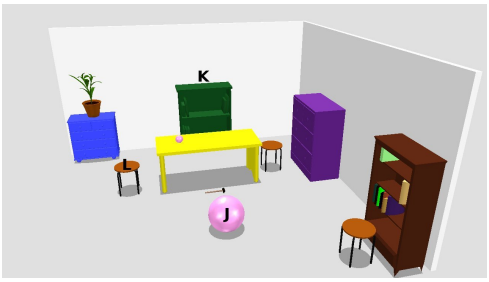
1.	*the plate next to the ball*	*perfect*>1
2.	*the large blue box*	*imperfect*>1
3.	*the red dish*	*perfect*=1
4.	*the brown bookcase under the portrait*	*imperfect*>1
5.	*the orange mug near the vase*	*imperfect*>1
6.	*the large plate*	*perfect*>1
7.	*the large green bookcase near the chest*	*imperfect*=1
8.	*the large ball on the table*	*imperfect*>1
9.	*the portrait above the bookcase*	*perfect*=1

(b) Descriptions with their characterization

Figure 2: Context visualization and object descriptions used in the Interpretive experiment.



(a) Projective relations and "end, edge, corner" and "center" of a table



(b) Colour, size, positional relation and intervening object in a room

Figure 3: Two of the scenarios used in the Generative experiments.

a description on the basis of what they understood (rather than using the designated referents). Each annotator handled half of the descriptions, and the other annotator verified the annotations. Disagreements were resolved by consensus.

Our study had 26 participants, who generated a total of 432 spoken descriptions (average length was 10 words, median 8, and the longest description had 21 words). We manually filtered out 32 descriptions that were broken up by the ASR due to pauses made by the speakers, and 105 descriptions that *Scusi? CantRepresent* (Section 6.2). Two sets of files were submitted to *Scusi?*: a set containing textual transcriptions of the remaining 295 descriptions, and a set containing textual alternatives produced by the ASR for each of these descriptions.

This experiment enables us to observe the frequencies of descriptions with different characteristics (Section 4), and determine their influence on performance, as well as the effect of ASR versus textual input. Table 3 displays the frequencies of the four accuracy classes of descriptions (*perfect* =1 and >1 and *imperfect* =1 and >1), and two knowledge classes (*known* and *unknown-OOV*) (Section 4). For instance, the top row shows that 197 descriptions are *known-perfect*=1 (Column 2), and 25 descriptions are *unknown-OOV* (Column 3). 18 *unknown-non-OOV* descriptions were omitted from Table 3. These descriptions have Gold ICGs, but contain word combinations that are not known to *Scusi?*, e.g., "on top of" and "at the front of". Note the low frequencies of three of the *unknown-OOV* categories, and of the *imperfect>1* classes. The latter suggests that, unlike our Interpretive trial, people rarely generate descriptions that are both ambiguous and inaccurate. Table 3 also displays the results obtained for the performance metrics *NotFound@K*, *FRecall@K* and *NDCG@K* (Section 6) for each accuracy-knowledge combination and for Text and ASR input; the results are described in Section 7.

6 Evaluation Metrics

We first consider the Interpretive trial followed by the Generative trial.

6.1 Interpretive trial

Scusi?'s understanding of each description was compared with that of our trial subjects by calculating the Spearman rank correlation coefficient and Pearson correlation coefficient between the average of the scores of the subjects' ratings for each object, and the probability assigned

Table 1: Descriptions that cannot be represented.

	Positional relation			Others and
	Poly-semous	Complex	No Landm.	Prep. Phrase "with"/"of"
perfect=1	9	29	0	9
perfect>1	5	15	0	4
imperfect=1	6	13	18	3
imperfect>1	2	2	1	0
TOTAL	22	59	19	16

by *Scusi?* to the top-ranked correct interpretation with the corresponding head object, e.g., `plate16-near-ball09` for the first description in Figure 2(b). The results for the Spearman rank and Pearson correlation coefficient appear in Section 7.1.

6.2 Generative trial

We first describe our broad metrics, followed by the fine-grained metrics.

CantRepresent counts the number of utterances that an SLU system cannot represent, which are a subset of the *unknown* utterances, and are excluded from the rest of the evaluation. Table 1 displays the frequencies of such descriptions and their causes (11 descriptions had more than one problem). As shown in Table 1, complex positional relations, e.g., "*the left front corner*"$*$, account for most of the problems.

NotFound@K counts the number of representable utterances for which no correct interpretation was found within rank K. *NotFound*@∞ considers all the interpretations returned by an SLU system. It is worth noting that *NotFound* utterances are included when calculating the following metrics.

Precision@K and Recall@K. The *@K* versions of precision and recall evaluate performance for different cut-off ranks K.

Precision@K is simply the number of correct interpretations at rank K or better divided by K. *Recall*@K is defined as follows:

$$\mathit{Recall@K}(d) = \frac{|\mathit{CF}(d) \cap \{I_1, \ldots, I_K\}|}{|C(d)|},$$

where $C(d)$ is the set of correct interpretations for utterance d, $\mathit{CF}(d)$ is the set of correct interpretations found by an SLU module, and I_j denotes an interpretation with rank j.

Contrary to IR settings, where typically there are many relevant documents, in language understanding situations, there is often one correct interpretation for an utterance (Table 3). If this interpretation is ranked close to the top, *Precision*@K will be constantly reduced as K increases. Hence, we eschew this measure when evaluating the performance of an SLU system.

An SLU module may return several equiprobable interpretations, some of which may be incorrect. The relative ranking of these interpretations is arbitrary, leading to non-deterministic values for *Recall*@K — a problem that is exacerbated when K falls within a set of such equiprobable interpretations. This motivates a variant of *Recall*@K, denoted *FRecall*@K (*Fractional Recall*), that allows us to represent the arbitrariness of the ranked order of equiprobable interpretations, as follows:

$$\mathit{FRecall@K}(d) = \frac{\sum_{j=1}^{K} \mathit{fc}(I_j)}{|C(d)|}, \qquad (1)$$

where *fc* is the fraction of correct interpretations among those with the same probability as I_j (this is a proxy for the probability that I_j is correct):

$$\mathit{fc}(I_j) = \frac{c_j}{h_j - l_j + 1}, \qquad (2)$$

where l_j is the lowest rank of all the interpretations with the same probability as I_j, h_j the highest rank, and c_j the number of correct interpretations between rank l_j and h_j inclusively.

Normalized Discounted Cumulative Gain (NDCG)@K. A shortcoming of *Recall*@K is that it considers the rank of an interpretation only in a coarse way (at the level of K). A finer-grained account of rank is provided by *NDCG*@K (Järvelin and Kekäläinen, 2002), which discounts interpretations with higher (worse) ranks.

DCG@K allows the definition of a relevance measure for a result, and divides this measure by a logarithmic penalty that reflects the rank of the result. Using $\mathit{fc}(I_j)$ as a measure of the relevance of interpretation I_j, we obtain

$$\mathit{DCG@K}(d) = \mathit{fc}(I_1) + \sum_{j=2}^{K} \frac{\mathit{fc}(I_j)}{\log_2 j}.$$

This score is normalized to the $[0, 1]$ range by dividing it by the score of an ideal answer where $|C(d)|$ correct interpretations are ranked in the first $|C(d)|$ places, yielding

$$\mathit{NDCG@K}(d) = \frac{\mathit{DCG@K}(d)}{1 + \sum_{j=2}^{\min\{|C(d)|, N\}} \frac{1}{\log_2 j}}. \qquad (3)$$

Note that *FRecall*@K is computed in relation to the number of correct interpretations, while *NDCG*@K considers the minimum of K and this number (Equations 1 and 3 respectively).

Table 2: Results of the Interpretive trials.

#	Survey	*Scusi?* I_1	*Scusi?* I_2	*Scusi?* I_3
1.	plate16	**plate16 –(near)→ ball9**	plate15 –(near)→ ball9	plate28 –(near)→ ball20
2.	box17	**box17**	box19	carpet23
3.	plate16	mug26	**plate16**	mug12
4.	bookcase14	bookcase10 –(under)→ portrait18	bookcase10 –(under)→ portrait8	***bookcase14*** –(instr_r)→ portrait8
5.	mug26	**mug26 –(near)→ vase6**	mug12 –(near)→ vase6	mug13 –(near)→ vase6
6.	plate28	plate16/**plate28**	**plate28**/plate16	plate15
7.	bookcase10	bookcase14 –(near)→ chest7	**bookcase10 –(near)→ chest7**	bookcase14 –(recipient_r)→ chest7
8.	ball9	**ball9 –(on)→ table1**	ball20 –(agent_r)→ table1	ball20 –(action_r)→ table1
9.	portrait18	**portrait18 –(above)→ bookcase10**	portrait8 –(above)→ bookcase10	portrait27 –(instr_r)→ bookcase14

7 Results

We first discuss the results of our Interpretive trials followed by those of our Generative trials.

7.1 Interpretive Trials

Table 2 compares the results of the Web survey with *Scusi?*'s performance for the Interpretive trials. Column 2 indicates the object preferred by the trial subjects, and Columns 3-5 show the top-three interpretations preferred by *Scusi?* (I_1–I_3). Matches between the system's output and the averaged participants' ratings are boldfaced.

As seen in Table 2, *Scusi?*'s ratings generally match those of our participants, achieving a strong Pearson correlation of 0.77, and a weaker Spearman correlation of 0.63. This is due to the fact that implausible interpretations get a score of 0 from *Scusi?*, while some people still choose them, thus yielding different ranks for them.

Scusi?'s top-ranked interpretation matches our participants' preferences in 5.5 cases, and its second-ranked interpretation in 2.5 cases (the fractions are for equiprobable interpretations). The discrepancies between *Scusi?*'s choices and those of our trial subjects are explained as follows: (desc. 3) "the red dish" – according to Leacock and Chodorow's similarity metric (Section 3.1), a mug is more similar to a dish than a dinner plate, while our trial subjects thought otherwise; (desc. 4) "the brown bookcase under the portrait" – *Scusi?* penalizes heavily attributes that do not match reality (Zukerman et al., 2008), hence `bookcase14` is penalized, as it is not under any portrait; (desc. 6) "the large plate" – our participants perceived `plate28` to be larger than `plate16` although they are the same size, and hence equiprobable; (desc. 7) "the large green bookcase near the chest" – like description 4, `bookcase10` (which is green) is ranked second due to its low probability of being considered large.

Thus, according to this trial, *Scusi?*'s performance satisfies our original requirement for reasonable behaviour and plausible mistakes, but perhaps it should be more forgiving with respect to mis-matched attributes.

7.2 Generative Trials

Table 3 displays the results for *NotFound*@K, *FRecall*@K and *NDCG*@K for $K = 1, 3, 10, \infty$ for Text and ASR input, the four accuracy classes, and the *known* and *unknown-OOV* knowledge categories. There are 277 descriptions in total (instead of 295), as 18 *unknown-non-OOV* descriptions were omitted from Table 3 (Section 5.2). As mentioned in Section 5.2, the vast majority of the utterances belong to the *perfect=1* class (with *known* or *unknown-OOV* words), and to the *known perfect>1* and *imperfect=1* categories.

ASR versus Text. The *NotFound*@1,3, *FRecall*@1,3 and *NDCG*@1,3 metrics show that *Scusi?* yields at least one correct interpretation at the lowest (best) ranks for the vast majority of Text inputs (the discrepancy between *FRecall* and *NDCG* at low ranks is due to the way these measures are calculated, Section 6.2). This suggests that in the absence of ASR errors, if correct interpretations are found, the system's confidence in its output is justified. As expected, the *NotFound* values are substantially higher, and the *FRecall* and *NDCG* values lower, for inputs obtained from the ASR (23% of the descriptions had one wrong word in the best ASR alternative, 21% had two wrong words, 12.5% had three, and 8.5% more than three). There is a substantial improvement in *FRecall* and *NDCG* as ranks increase, which shows that contextual information can alleviate some ASR errors. The improvement in these metrics for the *perfect>1* class, without affecting *NotFound*, indicates that *Scusi?* finds more correct interpretations for the same descriptions.

The ASR results compared to those of Text indicate that, unsurprisingly, speech recognition quality must be improved. This may be achieved through advances in ASR technology, or by pre-

Table 3: Description breakdown in terms of accuracy and knowledge, performance metrics and results.

	Known		**Unknown-OOV**	
	Text	ASR	Text	ASR
perfect=1	197		25	
NotFound@1,3,10,∞	9,4,2,1	73,60,49,31	8,8,8,3	16,13,11,9
FRecall@1,3,10,∞	0.95,0.98,0.99,0.99	0.61,0.69,0.75,0.84	0.47,0.68,0.68,0.88	0.24,0.45,0.54,0.64
NDCG@1,3,10,∞	0.95,0.98,0.98,0.98	0.61,0.69,0.71,0.73	0.47,0.64,0.64,0.68	0.24,0.40,0.44,0.46
perfect>1	30		1	
NotFound@1,3,10,∞	2,2,1,1	13,12,10,9	0,0,0,0	0,0,0,0
FRecall@1,3,10,∞	0.40,0.82,0.88,0.97	0.22,0.48,0.62,0.70	0.50,1.00,1.00,1.00	0.50,1.00,1.00,1.00
NDCG@1,3,10,∞	0.84,0.84,0.85,0.87	0.47,0.48,0.53,0.55	1.00,1.00,1.00,1.00	1.00,1.00,1.00,1.00
imperfect=1	18		2	
NotFound@1,3,10,∞	1,1,1,0	8,7,7,5	0,0,0,0	1,1,1,1
FRecall@1,3,10,∞	0.91,0.94,0.94,1.00	0.56,0.59,0.61,0.72	0.51,0.54,0.64,1.00	0.03,0.08,0.26,0.50
NDCG@1,3,10,∞	0.91,0.94,0.94,0.95	0.56,0.59,0.60,0.61	0.51,0.54,0.58,0.66	0.03,0.07,0.14,0.20
imperfect>1	3		1	
NotFound@1,3,10,∞	1,0,0,0	3,2,1,1	0,0,0,0	1,1,1,1
FRecall@1,3,10,∞	0.18,0.53,0.61,1.00	0.00,0.33,0.51,0.67	0.03,0.09,0.29,1.00	0.00,0.00,0.00,0.00
NDCG@1,3,10,∞	0.36,0.53,0.56,0.64	0.00,0.27,0.35,0.38	0.06,0.08,0.15,0.31	0.00,0.00,0.00,0.00

venting ASR errors (Gorniak and Roy, 2005; Sugiura et al., 2009) or correcting them (López-Cózar and Callejas, 2008; Kim et al., 2013).

Known versus Unknown-OOV. *Perfect=1* is the only class with a substantial number of OOV words (25). Note the increase in *FRecall* up to rank @∞ for *known* ASR and *unknown-OOV* Text and ASR, which indicates that correct interpretations are returned at very high ranks when input words are not identified (*NDCG* increases only modestly, as it penalizes high ranks). The difference in performance between *known-perfect=1* and *unknown-OOV-perfect=1* suggests that it is worth improving *Scusi?*'s vocabulary coverage.

8 Conclusion

We offered a framework for the evaluation of SLU systems that comprises a characterization of spoken utterances, experimental design and evaluation metrics. We described its application to the evaluation of *Scusi?*— our SLU module for the interpretation of descriptions in a household context.

Our characterization of descriptions identifies frequently occurring cases, such as *perfect=1*, and rare cases, such as *imperfect>1*; and highlights the influence of vocabulary coverage on performance.

Our two types of experiments enable the evaluation of an SLU system's performance from two viewpoints: Interpretive trials support the comparison of an SLU module's performance with that of people as addressees, and Generative trials assess the performance of an SLU system when interpreting descriptions commonly spoken by users. The results of the Interpretive trial were encouraging, but they indicate that *Scusi?*'s "punitive" attitude to attributes that do not match reality, such as a bookcase not being under any portrait, may need to be moderated. However, as stated above, *imperfect>1* descriptions were rare in our Generative trials. The results of these trials show that development effort should be invested in (1) ASR accuracy (Kim et al., 2013); (2) vocabulary coverage; and (3) ability to represent complex, polysemous and no-landmark positional relations. In contrast, descriptive prepositional phrases starting with "with" or "of" may be judiciously ignored, or the referent may be disambiguated by asking a clarification question.

Our *CantRepresent* and *NotFound* evaluation metrics provide an overall view of an SLU system's performance. IR-based metrics have been used in the evaluation of SLU systems to compare an interpretation returned by an SLU module with a reference interpretation. In contrast, we employ *FRecall* and *NDCG* in the traditional IR manner, i.e., to assess the rank of correct interpretations in an N-best list. The relevance measure *fc* (Equation 2), which is applied to both metrics, enables us to handle equiprobable interpretations. However, rank-based evaluation metrics do not consider the absolute quality of an interpretation, i.e., the top-ranked interpretation might be quite bad. In the future, we propose to investigate confidence/accuracy metrics, such ICE (Thomson et al., 2008), to address this problem.

Acknowledgments

This research was supported in part by grants DP110100500 and DP120100103 from the Australian Research Council.

References

A. Black, S. Burger, A. Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, J.D. Williams, K. Yu, S. Young, and M. Eskenazi. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the 11th SIGdial Conference on Discourse and Dialogue*, pages 2–7, Portland, Oregon.

CIE, 1995. *Industrial colour difference evaluation*. CIE 115-1995.

K.R. Coventry and S.C. Garrod. 2004. *Saying, Seeing, and Acting: the psychological semantics of spatial prepositions*. Psychology Press.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18(2):233–263.

D. DeVault, K. Sagae, and D. Traum. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the 10th SIGdial Conference on Discourse and Dialogue*, pages 11–20, London, United Kingdom.

C.D. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press.

S. Gandrabur, G. Foster, and G. Lapalme. 2006. Confidence estimation for NLP applications. *ACM Transactions on Speech and Language Processing*, 3(3):1–29.

P. Gorniak and D. Roy. 2005. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *ICMI'05: Proceedings of the 7th International Conference on Multimodal Interfaces*, pages 138–143, Trento, Italy.

L. Hirschman. 1998. The evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12:281–305.

K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

K. Jokinen and M. McTear. 2010. *Spoken Dialogue Systems*. Morgan and Claypool.

J.D. Kelleher and F.J. Costello. 2008. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

S.N. Kim, I. Zukerman, Th. Kleinbauer, and F. Zavareh. 2013. A noisy channel approach to error correction in spoken referring expressions. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan.

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–285. MIT Press.

R. López-Cózar and Z. Callejas. 2008. ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. *Journal of Speech Communication*, 50(8-9):745–766.

S. Möller. 2008. Evaluating interactions with spoken dialogue telephone services. In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, pages 69–100. Springer.

T. Paek. 2001. Empirical methods for evaluating dialog systems. In *SIGDIAL'01 – Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–9, Aalborg, Denmark.

G. Salton and M.J. McGill. 1983. *An Introduction to Modern Information Retrieval*. McGraw Hill, New York, New York.

J.F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA.

K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura. 2009. Bayesian learning of confidence measure function for generation of utterances and motions in object manipulation dialogue task. In *Proceedings of Interspeech 2009*, pages 2483–2486, Brighton, United Kingdom.

B. Thomson, K. Yu, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, and S. Young. 2008. Evaluating semantic-level confidence scores with multiple hypotheses. In *Proceedings of Interspeech 2008*, pages 1153–1156, Brisbane, Australia.

I. Zukerman, E. Makalic, M. Niemann, and S. George. 2008. A probabilistic approach to the interpretation of spoken utterances. In *PRICAI 2008 – Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence*, pages 581–592, Hanoi, Vietnam.