Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 154–185

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Employing distance-based semantics to interpret spoken referring expressions☆,☆☆

Ingrid Zukerman*, Su Nam Kim, Thomas Kleinbauer, Masud Moshtaghi

Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia

Received 26 May 2014; received in revised form 29 October 2014; accepted 12 January 2015
Available online 23 January 2015

Abstract

In this paper, we present *Scusi?*, an anytime numerical mechanism for the interpretation of spoken referring expressions. Our contributions are: (1) an anytime interpretation process that considers multiple alternatives at different interpretation stages (speech, syntax, semantics and pragmatics), which enables *Scusi?* to defer decisions to the end of the interpretation process; (2) a mechanism that combines scores associated with the output of the different interpretation stages, taking into account the uncertainty arising from a variety of sources, such as ambiguity or inaccuracy in a description, speech recognition errors and out-of-vocabulary terms; and (3) distance-based functions with probabilistic semantics that represent lexical similarity between objects' names and similarity between stated requirements and physical properties of objects (viz colour, size and positional relations). We considered two approaches for combining these descriptive attributes, viz multiplicative and additive, and determined whether prioritizing certain interpretation stages and descriptive attributes affects interpretation performance. We conducted two experiments to evaluate different aspects of *Scusi?*'s performance: Interpretive, where we compared *Scusi?*'s understanding of descriptions that are mainly ambiguous or inaccurate with people's understanding of these descriptions, and Generative, where we assessed *Scusi?*'s understanding of naturally occurring spoken descriptions. Our results show that *Scusi?*'s understanding of the descriptions in the Interpretive trial is comparable to that of people; and that its performance is encouraging when given arbitrary spoken descriptions in diverse scenarios, and excellent for the corresponding written descriptions. In both experiments, *Scusi?* significantly outperformed a baseline system that maintains only top same-score interpretations.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Spoken language understanding; Numerical approach; Semantic interpretation; Distance-based semantics; Performance evaluation

1. Introduction

People often express themselves ambiguously or inaccurately (Trafton et al., 2005; Moratz and Tenbrink, 2006; Funakoshi et al., 2012). An ambiguous reference to an object matches several objects well, while an inaccurate reference

☆ This paper has been recommended for acceptance by R.K. Moore.

☆☆ This paper describes research conducted on the *Scusi?* system, amending, extending and adding detail to the work of Zukerman et al. (2008), Makalic et al. (2008) and Kleinbauer et al. (2013).

* Corresponding author. Tel.: +61 399055202.
E-mail addresses: Ingrid.Zukerman@monash.edu (I. Zukerman), Su.Kim@monash.edu (S.N. Kim), Thomas.Kleinbauer@monash.edu (T. Kleinbauer), Masud.Moshtaghi@monash.edu (M. Moshtaghi).

http://dx.doi.org/10.1016/j.csl.2015.01.002
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

matches one or more objects partially. For instance, in a household domain, a reference to a "big blue mug" is ambiguous if there is more than one big blue mug in the room, and inaccurate if there are two mugs in the room, one big and red, and one small and blue. In addition, ambiguous or inaccurate references may result from different parse trees (e.g., due to variants in prepositional attachments), or from misheard utterances in spoken interactions. Computer systems that interact with people in natural language must be able to cope with such issues. This does not mean that a system must always obtain an intended interpretation of an utterance (although that would be nice), but when it misunderstands an utterance, the misunderstanding should be plausible.

Like Funakoshi et al. (2012), Lison and Kruijff (2009) and Ross (2010), we posit that simply considering the best interpretation of an utterance would not address these problems. In fact, our approach is part of a growing trend which harnesses numerical formalisms to consider multiple interpretations in order to handle the uncertainty inherent in real-world problems, e.g., (Funakoshi et al., 2012; Lison and Kruijff, 2009; Ross, 2010). However, we defer decisions to the end of the interpretation process, which like Punyakanok et al.'s (2008) approach, allows us to make global, rather than local, decisions. The score associated with an interpretation typically represents its goodness, which in turn enables the ranking of candidate interpretations. Additionally, scores may be used, in combination with utilities, to determine a course of action, e.g., ask a clarification question, or perform a requested action. Considering multiple interpretations, coupled with numerical scores, enables a system to recover from situations where the highest-scoring interpretation is not the intended one (e.g., due to speech recognition errors), or detect situations where several interpretations are similarly plausible.

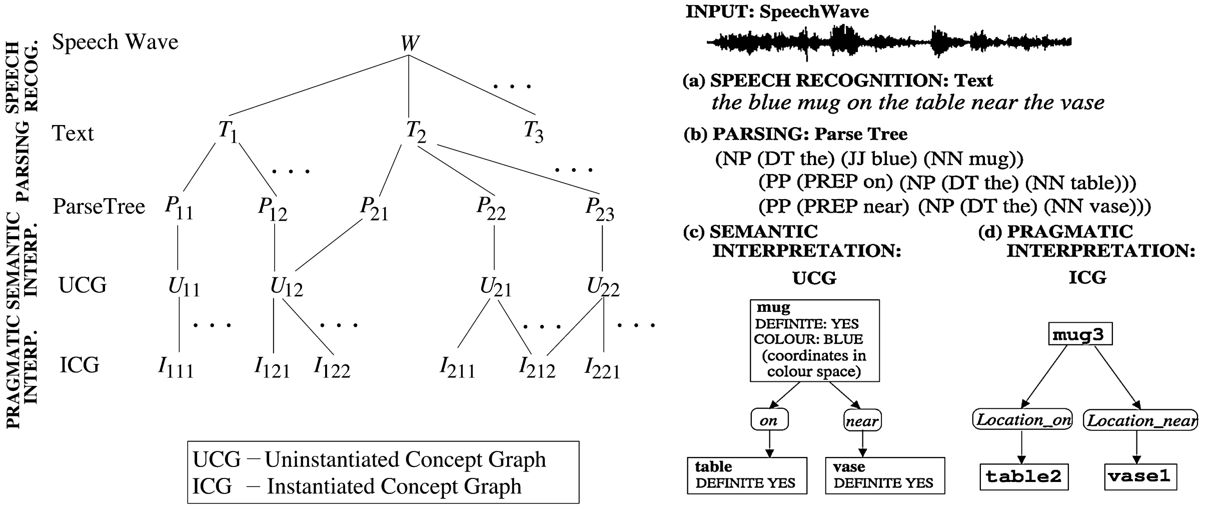
In this paper, we present a numerical mechanism for the interpretation of spoken referring expressions which considers multiple interpretations at different levels (i.e., speech, syntax, semantics and pragmatics), and incorporates the physical reality of the context at the pragmatics level of the interpretation process. The disparate processes carried out at each level of interpretation require our mechanism to combine scores and probabilities obtained from several sources. Specifically, our mechanism combines scores returned by an Automatic Speech Recognizer (ASR), probabilities estimated by a probabilistic parser, scores that estimate the complexity of an interpretation, and scores that represent how well the properties of candidate objects (viz lexical item, colour, size and location) match a user's requirements (Section 4). The scores may be viewed as *subjective probabilities*, which represent one's state of certainty regarding the truth of a proposition (Pearl, 1988).

Our mechanism produces a ranked list of interpretations at each of the above levels, culminating in instantiated objects in a specific context. For example, consider the description "the large blue mug" uttered in a room which contains a large aqua mug, a small blue mug and a large red mug, denoted `mug0`, `mug1` and `mug2` respectively, among other objects. Our mechanism yields alternative interpretations where in principle the referent may be each of the objects in the room, and each interpretation is associated with a score. Interpretations comprising `mug0`, `mug1` or `mug2` as referents have a higher score than interpretations comprising other objects (which have a low score); and interpretations where the referent is a blue object have a higher score than interpretations where the referent has another colour (this helps in situations where the name of a referent has been misheard, but its colour was heard correctly). Since none of the mugs in our example match the description precisely, their ranking depends on how well the given requirements match the actual colour and size of the mugs, and on the relative importance of colour and size. These rankings and their scores support the determination of a course of action by a robotic agent. For instance, if the score of, say, the `mug0` interpretation is significantly higher than the scores of the other mugs, the agent can simply retrieve `mug0`. If the scores of the interpretations involving the three mugs are similar, the agent should ask a clarification question. However, if the description was uttered in a room with no mugs, the scores of all the candidate objects would be low, which may prompt the agent to look for a mug in a different room. The implementation of such a decision process is the next step in this project.

Our mechanism was evaluated in two experimental settings: an *Interpretive* web-based setting to determine how well our system's understanding of given written referring expressions matches people's understanding; and the more common *Generative* setting, e.g., (Gandrabur et al., 2006; Thomson et al., 2008; DeVault et al., 2009), to assess our system's performance in various scenarios where people give spoken descriptions of designated referents.

The contributions of this paper are:

- An anytime interpretation process that considers multiple alternatives at different interpretation stages (speech, syntax, semantics and pragmatics), which enables our system to defer decisions to the end of the interpretation process.


SPEECH RECOG.
PARSING
SEMANTIC INTERP.
PRAGMATIC INTERP.
Speech Wave
Text
ParseTree
UCG
ICG
W
T1
T2
T3
P11
P12
P21
P22
P23
U11
U12
U21
U22
I111
I121
I122
I211
I212
I221
UCG – Uninstantiated Concept Graph
ICG – Instantiated Concept Graph
INPUT: SpeechWave
(a) SPEECH RECOGNITION: Text
the blue mug on the table near the vase
(b) PARSING: Parse Tree
(NP (DT the) (JJ blue) (NN mug))
(PP (PREP on) (NP (DT the) (NN table)))
(PP (PREP near) (NP (DT the) (NN vase)))
(c) SEMANTIC INTERPRETATION:
UCG
mug
DEFINITE: YES
COLOUR: BLUE
(coordinates in colour space)
on
near
table
DEFINITE YES
vase
DEFINITE YES
(d) PRAGMATIC INTERPRETATION:
ICG
mug3
Location_on
Location_near
table2
vase1

(a) Stages of the interpretation process

(b) Sample structures for the interpretation stages

Fig. 1. *Scusi?*'s spoken language interpretation process.

- A mechanism that combines the probabilities and scores associated with the output of the different interpretation stages, taking into account the uncertainty arising from a variety of sources, such as ambiguity or inaccuracy in a description, speech recognition errors and out-of-vocabulary terms (Section 4).
- Distance-based functions with probabilistic semantics that represent lexical similarity between objects' names and similarity between stated requirements and physical properties of objects, viz colour, size and topological and projective positional relations (Section 4).
- Two approaches for combining these distance-based functions, viz multiplicative and additive (Section 4); and experiments that determine whether prioritizing certain interpretation stages and descriptive attributes affects interpretation performance (Section 5.2).
- Evaluation with two corpora that illustrate a variety of conditions, which, to the best of our knowledge, are more diverse than those examined to date (Moratz and Tenbrink, 2006; Funakoshi et al., 2012; Kelleher and Costello, 2008) (Section 5).

This paper is organized as follows. Section 2 presents the interpretation process, and Section 3 details the semantic interpretation procedure. The estimation of the score of an interpretation appears in Section 4. Our evaluation experiments are described in Section 5. Related research and concluding remarks are given in Sections 6 and 7 respectively.

2. Searching for an interpretation

Our interpretation mechanism is implemented in a module called *Scusi?*, which processes spoken input in four stages: speech recognition, parsing, semantic interpretation and pragmatic interpretation (Fig. 1(a)). In the first stage of our interpretation process, *Scusi?* runs Automatic Speech Recognition (ASR) software (Microsoft Speech SDK 6.1) to generate candidate texts from a speech signal (up to 50 texts are produced). Each text is assigned a score that reflects the certainty of the ASR regarding the words in the text given the speech wave. The second stage applies Charniak's probabilistic parser (http://bllip.cs.brown.edu/resources.shtml#software) to generate parse trees from the texts. The parser generates up to 50 parse trees for each text, associating each parse tree with a probability. During semantic and pragmatic interpretation, parse trees are successively mapped into two representations based on Conceptual Graphs (Sowa et al., 1984): *Uninstantiated Concept Graphs (UCGs)* and *Instantiated Concept Graphs (ICGs)*. UCGs are obtained from parse trees deterministically – one parse tree generates one UCG (but a UCG may have several parent parse trees), and each UCG can generate many ICGs (an ICG may have several parent UCGs, which are obtained from similar parse trees).

A UCG represents syntactic information pertaining to the head nouns in the parent parse tree and their modifiers, and relations between concepts, which are directly derived from syntactic information in the parse tree and prepositions. A UCG also stores semantic information about head-noun modifiers whose meaning is known to *Scusi?*, viz colour and size.[1] Fig. 1(b) illustrates the generation of a candidate UCG and a matching ICG for the description "the blue mug on the table near the vase", where, as shown in the parse tree, the mug is expected to be near the vase (an alternative interpretation would place the table near the vase). The nouns "mug", "table" and "vase" in the parse tree are mapped to the concepts **mug**, **table** and **vase** in the UCG respectively, with the mug having a COLOUR attribute (represented by coordinates in the CIE94 colour space, Appendix A). The prepositions "on" and "near" in the parse tree are respectively mapped to the relations *on* and *near* in the UCG. In order to favour simple interpretations, the score associated with a UCG is the reciprocal of the number of nodes in it.

An ICG represents pragmatic information comprising instantiated concepts and relations between these concepts in the current context. Instantiated concepts are objects in the physical context (e.g., a particular room), and instantiated relations correspond to positional relations. The objects and positional relations within an ICG mirror those within its parent UCG(s), but in contrast to the generic names of UCG concepts (e.g., **mug**), an ICG contains specific objects (e.g., mug3, which is a candidate match for **mug**, COLOUR: BLUE in the example in Fig. 1(b)). The score of an ICG reflects how well its objects and relations match the specifications represented in its parent UCG(s) and the context. For instance, the score of the ICG in Fig. 1(b) reflects how well mug3, table2 and vase1 match the concepts **mug**, **table** and **vase** in the UCG respectively (a cup yields a worse match with the concept **mug** than a mug), how well the colour of mug3 matches the specified colour blue, whether mug3 is on table2, and how close it is to vase1.

The generation of ICGs is described in Section 3. In the mean time, it is worth noting that when generating pragmatic interpretations, the first step consists of proposing candidate ICG objects and positional relations for each UCG concept and relation respectively. *Scusi?* then checks whether a proposed combination of objects and positional relations satisfies the specifications in the UCG and matches the context. So, all the mugs and cups in the room are good candidates for the concept **mug** (with blue mugs being better candidates), all the tables and desks are good candidates for **table**, and all the vases for **vase**. When the relation *Location_on* is tested for a particular mug–table combination, it will produce a good match only for the mugs that are on tables, and when the relation *Location_near* is tested for a mug–vase combination, only the mugs that are near vases will be winners. So, if mug3 in Fig. 1(b) is blue, it is a good candidate in terms of lexical match and colour, but if it is not on table2 or near vase1, then the ICG in Fig. 1(b) will receive a low score. In contrast, if mug3 was on table1, the ICG representing this situation would have a better score.

The consideration of all possible options at each stage of the interpretation process is computationally intractable. Hence, *Scusi?* employs an *anytime algorithm* combined with a *thresholding approach* to generate interpretations in real time, and a *stochastic optimization strategy* to avoid getting stuck in local maxima.

The **anytime algorithm** ensures that the system can return a list of ranked interpretations at any point in time (after a start-up phase). At each stage of the interpretation process, the algorithm applies a selection–expansion cycle to add an element to a search graph as follows (Fig. 1(a)). First, it probabilistically selects a level to be expanded (speech, text, parse tree or UCG), and then it selects a node in the search graph at this level. For instance, if the speech level is selected, the next text produced by the ASR (in descending order of score) is returned. If the text level is selected, the algorithm chooses a text node to be expanded, and returns the next parse tree for it. This text node may be one that has produced no children yet (e.g., Text T_3 in Fig. 1(a)), or one that has already generated children (e.g., T_1). When a node is expanded for the first time, a priority buffer containing at most $k_{\max}$ ranked children is created, but only a single (top-ranked) child from this buffer is incorporated in the search graph. Every time a node is expanded, the next child from its buffer is added to the search graph. For example, when UCG U_{12} in Fig. 1(a) is expanded for the second time, the ICG-generation module returns the next ICG in the UCG's buffer, i.e., I_{122} (Section 3). The selection–expansion process is repeated until one of the following happens: all options are fully expanded, a time limit is reached, or a specific number of iterations is performed, which is the option chosen for the experiments described in Section 5. At any point after completing an expansion, the anytime algorithm can return a list of ranked ICGs with their ancestor sub-interpretations (text, parse tree(s) and UCG(s)).

[1] Other noun modifiers are simply denoted by the label ATTRIBUTE. Although at present these modifiers are not meaningful to *Scusi?*, the system can often overcome this limitation by taking advantage of redundant information in a description (Section 5). Additional attributes, such as shape and texture, may be incorporated into our framework by adding suitable information to *Scusi?*'s knowledge base.

To encourage the early generation of complete interpretations, when selecting a level to be expanded, preference is given to later stages in the search. That is, the probability of expanding the UCG level is higher than that of expanding the parse tree level, which in turn is higher than the probability of expanding the Text level, and finally the Speech Wave level. The **stochastic optimization strategy** is employed within a level during node selection: the most promising node is chosen most often, but not always. Specifically, nodes with a proven "track record" are preferred, i.e., nodes that have previously produced children with high probabilities/scores.

The **thresholding approach** comprises two efficiency measures. The first one, inspired by the branch and bound algorithm, does not expand sub-interpretations whose scores are significantly worse than the score of the top *N*th ICG, as they are unlikely to yield ICGs that are better than the current top *N* ICGs (*N* is currently set to 10). The second efficiency measure is based on the observation that the scores of the texts returned by the ASR drop significantly after a certain number of texts, as do the probabilities of the parse trees returned by the probabilistic parser. We harness this observation to prevent the procreation of unpromising nodes at all levels as follows. When the probability or score of the next child of a parent node *n* drops below a particular threshold relative to the probability or score of the best child of *n*, no additional children of *n* are generated. For example, given an ASR output threshold of 0.6, if the score of a textual output for the speech wave is less than 60% of the score of the first (best) textual output, then this text is ignored, and no more texts are generated for the speech wave. At present, we employ the following thresholds, which were set manually: 0.45 for ASR output, 0.1 for parse trees, and 0 for UCGs and ICGs (a parse tree generates only one UCG, and all ICGs are considered). Clearly, the values of these parameters can be learned in the same way we learn weights for the stages of the interpretation process and descriptive attributes (Section 5.2) – a task we propose to undertake in the future.

3. Generating semantic and pragmatic interpretations

UCGs are generated from parse trees in a rule-based manner as follows: parse tree nodes corresponding to head nouns are converted to UCG concepts; prepositions and syntactic relations are converted to UCG relations; and noun modifiers are incorporated as attributes of the concepts corresponding to head nouns (Fig. 1(b)).[2]

The process of generating ICGs from a UCG and computing their score is carried out by Algorithm 1, which has two main stages: *concept and relation postulation* (Steps 2–8) and *ICG construction* (Steps 9–14). This algorithm generates a buffer containing up to $k_{\max}$ ICGs ranked in descending order of score the first time a UCG is expanded. Every time a new ICG is requested for that UCG, the next-ranked ICG is returned.

Algorithm 1. Generate candidate ICGs for a UCG

```
Require: UCG $U$ comprising concepts and relations, context $\mathcal{C}$
1: Initialize buffer $\mathcal{I}_U$ of size $k_{\max}$
{Postulate objects and relations for the ICG}
2: for all concepts and relations $u$ in $U$ do
3:     Initialize a list of candidate concepts and relations $L_u \leftarrow \emptyset$
4:     for all instantiated objects and relations $k$ in the knowledge base do
5:         Compare the requirements in $u$ with the intrinsic attributes of $k$, yielding a score for the match
6:         Insert $k$ in the list $L_u$ in descending order of score
7:     end for
8: end for
{Construct ICGs}
9: for $j = 1$ to $k_{\max}$ do
10:    Generate the "next best" ICG $I_j$ by going down each list $L_u$ in turn
11:    Calculate the positional scores of the object-relation-object trigrams in ICG $I_j$
12:    Combine the scores obtained in Step 5 with the scores obtained in Step 11 (Eqs. (3) and (4), Section 4)
13:    Insert $I_j$ into buffer $\mathcal{I}_U$ in descending order of score
14: end for
```

[2] In the future, we will investigate an approach based on a chunking parser, similar to that of Meena et al. (2012), for the generation of UCGs.

Table 1
Concepts and relations used to build ICGs for the description “the blue mug on the table near the vase”.

mug, COLOUR: BLUE	**table**	**vase**	*on*	*near*
mug3	table0	vase0	*Location_on*	*Location_near*
cup0	table1	vase1	*Origin_on*	*Location_on*
mug0	table2	table0	*Destination_on*	*Origin_on*
cup1	mug0	table1	*Location_in*	*Destination_on*
table0	mug3	mug0	*Location_near*	*Location_in*
vase1	cup0	mug3	...	...
...	...	...		

3.1. Postulating concepts and relations

In this stage, the algorithm consults *Scusi?*’s knowledge base to propose a list of instantiated nodes (concepts and relations) for each UCG node; this list is sorted in descending order of score, which reflects how well the instantiated nodes match the corresponding node in the parent UCG. The objects in the knowledge base, which define the context, are represented by means of the terms commonly used to refer to them, their colour and dimensions, and their position in a three-dimensional space, e.g., a room. Relations are represented by their reference terms and by procedures that implement their operational semantics, e.g., what does it mean in the physical world for an object to be near or on a landmark?

In **Step 5**, for each concept (relation) u in UCG U, we estimate the match score between each instantiated concept (relation) k in *Scusi?*’s knowledge base and u. To this effect, we compare the intrinsic attributes of each instantiated object k (i.e., lexical item, colour and size) with the requirements stated in concept u, and the lexical item of each instantiated relation with the term used for relation u. For instance, given the UCG concept **cup**, the instantiated concepts {cup0, ..., cup5} have a perfect lexical match with the UCG concept, while {mug0, ..., mug4} have a good lexical match. If the UCG concept also had a COLOUR: BLUE attribute, then the colour coordinates of the objects in the knowledge base (including mugs and cups) would be matched against the colour coordinates for BLUE (as indicated in Section 1, matches with non-cup-like concepts are useful when the head noun has been misheard, but there are objects that match the specified colour). The score for lexical, colour and size match is calculated using the functions outlined in Table 2 and described in Appendix A.

This stage of the algorithm yields a list of candidate instantiated concepts (relations) L_u sorted in descending order of score for each UCG concept (relation) u. Table 1 shows the top objects in the sorted lists generated for the description in Fig. 1(b) “the blue mug on the table near the vase” in a context that comprises cup0 and mug3 (which are blue), and cup1 and mug2 (which are not blue), three tables, two vases and other objects (the reasonable candidates appear in larger font). The blue mug is ranked first, followed by the blue cup, the other mug and cup, and then all the other objects in the scene (which have a very low match score with the lexical item “mug”, with blue objects preceding objects of other colours). The three tables have a perfect lexical match with the second noun in the description, hence they have the same score (followed by vases, mugs, cups and other objects – all with a very low lexical match score). Similarly, the two vases have a perfect lexical match with the third noun (followed by tables, mugs, cups and other objects). There are three equal-score relations which match the preposition *on*, and one relation that matches *near*, with other relations yielding poor lexical matches with these prepositions.

3.2. Constructing ICGs

In this stage, the algorithm uses the list of instantiated concepts (relations) built for each concept (relation) in a UCG to construct candidate ICGs for this UCG, and sorts these ICGs in descending order of score. First, **Step 10** applies an enumerative process to generate different combinations of concepts and relations from the L_u lists maintained for the UCG nodes. This is done by iteratively selecting one candidate instantiated concept (relation) from each list starting at the top. For instance, the concepts and relations in Table 1 are combined as follows to build candidate ICGs. First, the top line {mug3, table0, vase0, *Location_on*, *Location_near*}, which has the highest total score, is used.

Table 2
Summary of functions for calculating the scores of intrinsic attributes and positional relations.

Function	Calculation method
Intrinsic scores	
Sc($u_{\text{lex}}\|k$)	Linear function based on Leacock and Chodorow's WordNet distance metric (Leacock and Chodorow, 1998); compares the lexical item in UCG node u (e.g., "mug") with a canonical term used to refer to a candidate instantiated concept (e.g., "cup" for `cup0`)
Sc($u_{\text{col}}\|k$)	Linear function based on distance in the CIE94 (L, a, b) colour space (Rubner et al., 2001); compares the colour coordinates of the colour mentioned in a description (e.g., "pink") with the colour coordinates of a candidate referent

Sc($u_{\text{size}} \in \text{big}\|u_{\text{lex}}, k$)
Sc($u_{\text{size}} \in \text{small}\|u_{\text{lex}}, k$)
	Sigmoid functions around the mean size or dimension of objects of the same type as k that are deemed big and objects that are deemed small; for instance, given a requested size or dimension, such as "big mug" or "short bookcase", compares the size or dimension of a candidate object, e.g., `mug3` or `bookcase4`, to the mean size of big mugs or the mean height of small bookcases respectively

Positional scores: topological relations	
Sc($on(k_r, k_l)$)	Linear function of the overlap between k_r and k_l in the horizontal xy plane, provided the bottom of k_r touches the top of k_l; *on* in the vertical plane (e.g., "on the wall") is a rotated version of this function
Sc($in(k_r, k_l)$)	Linear function of the volume of k_r that is within k_l
Sc($near(k_r, k_l)$)	Base 2 negative exponent function of the distance between k_r and k_l and the maximum of their surface area, so that small objects must be quite close to be considered *near* each other, while bigger objects can be farther apart
Sc($at(k_r, k_l)$)	Combination of *on*, *in* and *near*
Sc($Fn(k_r, k_l)$)	For $Fn \in$ {*in/at/near the center of*}, *Fn* is implemented as a combination of *on* and *near*, where the center of the landmark k_l is a small transparent square at the center of k_l; for $Fn \in$ {*in/at/near the corner/edge/end of*}, *Fn* is implemented as *near* the corner/edge/end (either *on* or *off* the landmark), where a corner is a small transparent square at each corner of the landmark k_l, and an edge/end is a ruler overlaying each end/edge of k_l
Positional scores: projective relations	
Sc($Fn(k_r, k_l)$)	For $Fn \in$ {*in front/back of*, *behind*, *to the left/right of*}, these relations, except *behind*, may take place *on* or *off* the landmark (an object can be *behind* the landmark only if it is *off* the landmark); for the *on* option, *Fn* is implemented as a combination of *near* and *on*, e.g., "the ball in front of the table" should be *near* the front of the table and *on* the table; for the *off* option, *Fn* demands a particular position of the candidate object relative to the landmark and the speaker, and a non-null projection of the object onto an appropriate plane, e.g., a candidate ball for "the ball in front of the table" must be located between the table and the speaker, and project onto a vertical plane facing the speaker
Sc($above(k_r, k_l)$)	Similar to *on*, but the object and the landmark do not have to touch; in fact, they do not even have to be *near* each other
Sc($under(k_r, k_l)$)	Similar to *above*, but the object must be under the *top* of a hollow landmark, e.g., a table, or fully under a convex landmark

The next two combinations are generated by replacing `table0` with `table1` and `table2`, as the tables have the same score, yielding {`mug3`, `table1`, `vase0`, *Location_on*, *Location_near*} and {`mug3`, `table2`, `vase0`, *Location_on*, *Location_near*}; and so on.

The scores of the object–relation–object trigrams in each ICG are estimated in **Step 11**. These scores reflect the extent to which the relationships between neighbouring object nodes in an ICG match the context (Section 4). For example, given the referring expression "the cup on the table", if, according to the knowledge base, the coordinates of `mug3` place it on `table2`, an ICG that contains [`mug3` – *Location_on* – `table2`] is assigned a high score for positional match; whereas if the coordinates of `cup1` place it on a shelf, the ICGs containing `cup1` are assigned low scores for positional match (the calculation of the scores for positional relations is outlined in Table 2 and described in Appendix B). In **Step 12**, these positional scores are combined with the scores calculated in Step 5 to obtain the total score of an ICG produced for a given UCG. If an ICG has more than one parent UCG, the scores obtained from the different parent UCGs are combined (Section 4.1.1). The resultant ICG is inserted in a buffer for its parent UCGs, which is sorted in descending order of the ICGs' scores.

4. Estimating the score of an interpretation

As mentioned above, the score of an ICG may be viewed as a subjective probability representing *Scusi?*'s state of certainty regarding the match between a candidate ICG and the intended meaning of a description. *Scusi?* ranks candidate ICGs according to this score, which is estimated on the basis of how well an ICG matches a heard description and the contextual information. At present, the context comprises the type, colour, dimensions and position of the objects in a room. This information could be provided by a scene analysis system, but it is currently obtained from a virtual representation generated with the `SweetHome` software package (http://sweethome3d.com). In the future, the context will also take into account salience from dialogue history.

Given a speech signal S and a context $\mathcal{C}$, $\mathrm{Sc}(I|S, \mathcal{C})$, the score of an ICG I, is estimated as follows:

$$\mathrm{Sc}(I|S, \mathcal{C}) \propto \mathrm{Sc}(I|U, \mathcal{C})^{W_I}\,\mathrm{Sc}(U|P)^{W_U}\,\mathrm{Pr}(P|T)^{W_P}\,\mathrm{Sc}(T|S)^{W_T} \tag{1}$$

where a UCG, parse tree and plain-text interpretation are denoted by U, P and T respectively, and the probability returned by the parser and the scores produced by the ICG and UCG stages and the ASR are each assigned a weight to adjust their effect on the final score of ICG I. These weights were determined in a modest learning experiment performed on a small development corpus, where we compared *Scusi?*'s performance using weights learned by two irrevocable search algorithms, viz a genetic algorithm and steepest ascent hill climbing, with that of a baseline were all the weights equal 1 (Section 5.2).

The ASR returns an estimate for $\mathrm{Sc}(T|S)$ which is in the [0, 1] range, and the parser returns the probability $\mathrm{Pr}(P|T)$. To estimate $\mathrm{Sc}(U|P)$ and $\mathrm{Sc}(I|U, \mathcal{C})$ we adopt an approach inspired by the Minimum Message Length (MML) principle (Wallace, 2005), which balances model complexity against data fit. MML operationalizes Occam's Razor, which may be stated as follows: "If you have two theories, both of which explain the observed facts, you should use the simplest until more evidence comes along". Thus, MML favours simple models that explain the data well. In our case, these are simple Concept Graphs that match a description and the context well. We model Concept Graph simplicity by penalizing UCGs with a high number of nodes as follows (the conversion from parse trees to UCGs is deterministic, so no other factor needs to be considered):

$$\mathrm{Sc}(U|P) = \frac{1}{N} \tag{2}$$

where N is the number of nodes.

$\mathrm{Sc}(I|U, \mathcal{C})$ represents data fit, i.e., how well an ICG I matches its parent UCG U and the context $\mathcal{C}$. These scores are mapped to the [0, 1] range to have a compatible range with that of the scores obtained for UCGs and texts and the probabilities returned by the parser. This compatible range in turn facilitates the weight-learning process described in Section 5.2.

We distinguish between two types of features that influence $\mathrm{Sc}(I|U, \mathcal{C})$: *intrinsic* and *positional*.

- ***Intrinsic features*** represent the attributes of individual nodes in an ICG. They determine how well an object (obtained from a room or context) or a relation in a candidate ICG matches the specifications in the corresponding concept or relation in its parent UCG (recall that a UCG concept node is composed of a head noun and its modifiers, and a relation node is built from a preposition or prepositional expression, e.g., "to the left of"). Each object in *Scusi?*'s knowledge base has a type, colour (represented by CIE94 coordinates (Rubner et al., 2001)) and size (represented by dimensions of the bounding box circumscribing the object (Skubic et al., 2004)). Hence, the intrinsic features handled by *Scusi?* at present are: *lexical item*, *colour* and *size* (Appendix A). In addition, *Scusi?* identifies *unknown* attributes, which contain noun modifiers that *Scusi?* does not understand. For instance, if the description in Fig. 1 had been "the blue *ceramic* mug on the table near the vase", "ceramic" would have been an unknown attribute, which would have reduced the score of all the matches with candidate objects. It is worth noting that when unknown attributes are due to ASR error, this behaviour affects the interpretations generated from the incorrect text.
- ***Positional features*** represent the relative location of objects. They model how well each object–position–object triple in an ICG matches the current context. For instance, given the description "the blue mug *on* the table *near* the vase" in Fig. 1(b), the relation `mug3`–*Location_on*–`table2` in the candidate ICG matches the context if the z coordinate of the bottom of `mug3` is the same as the z coordinate of the top of `table2`. Now, if `mug3` is also near `vase1`, then both positional relations in the ICG in Fig. 1(b) will match the context. However, if `mug3` is not near `vase1` or any vase, but `table2` is near, say, `vase3`, then the following ICG (obtained from a different parse tree and UCG) would yield the best match: `mug3`–*Location_on*–`table2`–*Location_near*–`vase3`. At present, *Scusi?* handles positional relations represented by the topological prepositions *on, in, near, at* and *in/at/near the center/corner/edge/end of*, which designate a region relative to a landmark; and positional relations represented by

the projective prepositions *in front/back of*, *behind*, *to the left/right of, above* and *under*, which describe a region projected from the landmark in a particular direction (Appendix B). The specification of the direction depends on the frame of reference being used (absolute, intrinsic or viewer centered) (Kelleher and Costello, 2008; Coventry and Garrod, 2004; Liu et al., 2010). In addition, projective relations depend on whether the landmark has a "face" (e.g., TV) or is human centric (e.g., chair) (Liu et al., 2010). In these cases, there is a high potential for ambiguity between the viewer-centered and intrinsic frames of reference. The consideration of these factors is left for future work, and like Kelleher and Costello (2008), in this research we adopt a viewer-centered frame of reference for projective relations.

Coventry and Garrod (2004) propose three factors that influence the use of positional prepositions: (1) the geometry of the situation, (2) control and support (e.g., a referent is *in* a landmark if moving the landmark controls the movement of the referent), and (3) domain knowledge (about the manner in which the referent and the landmark interact). In this paper, we consider only the first factor, as positional information may be obtained from a scene analysis system, while the extra-geometric information relies on additional user experiences. The representation and acquisition of this information and its incorporation into our geometric formalism is the subject of future work.

We consider two schemes for estimating the score of an ICG on the basis of intrinsic and positional features: *Multiplicative* and *Additive*. This score is later mapped to the [0, 1] range as described in Section 4.1.

Multiplicative scheme. This scheme is similar to that used in Eq. (1):

$$\begin{aligned}\mathrm{Sc}_{\mathrm{MULT}}(I|U,\mathcal{C}) = &\prod_{i=1}^{M}\prod_{j=1}^{M}\mathrm{Sc}(\mathrm{loc}(k_i,k_j))^{W_{\mathrm{loc}}\delta(\mathrm{loc}(u_i,u_j))}\\ &\times\prod_{i=1}^{M}\Big\{\mathrm{Sc}(u_{i,\mathrm{lex}}|k_i)^{W_{\mathrm{lex}}}\quad\times\quad\mathrm{Sc}(u_{i,\mathrm{col}}|k_i)^{W_{\mathrm{col}}\delta(u_i,\mathrm{col})}\times\\ &\qquad\mathrm{Sc}(u_{i,\mathrm{size}}|u_{i,\mathrm{lex}},k_i)^{W_{\mathrm{size}}\delta(u_i,\mathrm{size})}\times\mathrm{Sc}(u_{i,\mathrm{unk}})^{W_{\mathrm{unk}}\delta(u_i,\mathrm{unk})}\Big\}\end{aligned} \tag{3}$$

where

- M is the number of objects in ICG I.
- $\mathrm{Sc}(u_{i,\mathrm{lex}}|k_i)$ denotes the lexical similarity between the term $u_{i,\mathrm{lex}}$ and terms commonly used to refer to object k_i; $\mathrm{Sc}(u_{i,\mathrm{col}}|k_i)$ denotes the visual similarity between the term $u_{i,\mathrm{col}}$ and the colour of object k_i; $\mathrm{Sc}(u_{i,\mathrm{size}}|u_{i,\mathrm{lex}},k_i)$ denotes how well the term $u_{i,\mathrm{size}}$ reflects the size of object k_i; $\mathrm{Sc}(u_{i,\mathrm{unk}})$ is the penalty associated with a term that *Scusi?* does not understand (currently set to a very small value ϵ); and $\mathrm{Sc}(\mathrm{loc}(k_i,\ k_j))$ reflects how well the location of referent k_i matches location loc with respect to landmark k_j. Note that size is conditioned on lexical item, because size depends on both the described concept and the candidate object. For example, given a request for a large desk, a candidate table should be deemed large compared to desks, rather than compared to tables. In contrast, if a red table is requested, we only need to determine how close to red is the colour of a candidate table.
- $\delta(u_i, f)$ equals 1 if feature f was specified for object i, and 0 otherwise; similarly, $\delta(\mathrm{loc}(u_i,\ u_j))$ equals 1 if relation loc was specified between objects k_i and k_j, and 0 otherwise.
- the weights W_{lex}, W_{col}, W_{size}, W_{unk} and W_{loc} reflect the influence of lexical item, colour, size, unknown attribute and location respectively on the score of an interpretation. Like the weights assigned to the different stages of the interpretation process, these weights were experimentally determined (Section 5.2).

The second and third lines in Eq. (3) represent how well each object k_i in ICG I matches the lexical item, colour and size specified in its parent concept u_i in UCG U, and whether u_i has unknown attributes that could not be matched. The first line represents how well the relative location of two objects k_i and k_j in context $\mathcal{C}$ (e.g., a room) matches their specified location in UCG U (e.g., $on(k_i,\ k_j)$). For instance, given the ICG in Fig. 1(b), the second line in Eq. (3) estimates how suitable the term "mug" is for designating `mug3` and how well "blue" matches the colour of `mug3` (no size was specified), and how suitable the terms "table" and "vase" are for designating `table2` and `vase1` respectively (no colour or size was specified). The first line estimates to what extent mug3 could be said to be on table2 and near vase1; if the mug is on the table but far from the vase, the first score will be high and the second one low.

This scheme is rather unforgiving of partial matches or mismatches, e.g., the score of a lexical match between “mug” and cup1, which is less than 1, is substantially reduced when raised to an exponent greater than 1; and a mismatch of a single attribute in an ICG significantly lowers the score of the ICG. This motivates the more forgiving Additive scheme.

Additive scheme. This scheme estimates $\mathrm{Sc}_{\mathrm{ADD}}(I|U, \mathcal{C})$ using the following formulation:

$$\begin{aligned}\mathrm{Sc}_{\mathrm{ADD}}(I|U, \mathcal{C}) = &\sum_{i=1}^{M}\sum_{j=1}^{M}\mathrm{Sc}(\mathrm{loc}(k_i, k_j))W_{\mathrm{loc}}\delta(\mathrm{loc}(u_i, u_j)) \\ &+\sum_{i=1}^{M}\left\{\mathrm{Sc}(u_{i,\mathrm{lex}}|k_i)W_{\mathrm{lex}} + \mathrm{Sc}(u_{i,\mathrm{col}}|k_i)W_{\mathrm{col}}\delta(u_i, \mathrm{col})+\right. \\ &\qquad\left.\mathrm{Sc}(u_{i,\mathrm{size}}|u_{i,\mathrm{lex}}, k_i)W_{\mathrm{size}}\delta(u_i, \mathrm{size}) + \mathrm{Sc}(u_{i,\mathrm{unk}})W_{\mathrm{unk}}\delta(u_i, \mathrm{unk})\right\}\end{aligned} \tag{4}$$

4.1. *Score adjustment and normalization*

The anytime and stochastic operation of our algorithm, and the diverse sources from which the probabilities or scores are obtained, highlight two issues that must be addressed when calculating the score of an interpretation: (1) multiple parent nodes, and (2) probabilities or scores of different magnitudes.

4.1.1. *Multiple parent nodes*

As shown in Fig. 1(a), an ICG may have more than one parent UCG, which in turn may have more than one parent parse tree. To take into account multiple parents when estimating the score of ICGs and UCGs, we employ a normalized weighted average of the score of the child node as shown in Eqs. (5) and (6), where the weights are the probabilities or scores of the parent nodes.

$$\mathrm{Sc}(U_i|P) \propto \sum_j \mathrm{Sc}(U_i|P_j)\cdot \Pr(P_j), \quad \text{where } P_j \text{ is a parent parse tree of } U_i \tag{5}$$

$$\mathrm{Sc}(I_i|U, \mathcal{C}) \propto \sum_j \mathrm{Sc}(I_i|U_j, \mathcal{C})\cdot \mathrm{Sc}(U_j), \quad \text{where } U_j \text{ is a parent UCG of } I_i \tag{6}$$

The resultant scores are then mapped to the [0, 1] range.

4.1.2. *Probabilities and scores from diverse sources*

There are large variations in the probabilities and scores returned by the different interpretation stages. In particular, the probabilities returned by the parser are several orders of magnitude smaller than the scores returned by the other stages. In order to obtain probabilities and scores of a similar magnitude, we adopt two approaches: (1) adjusting the probabilities returned by the parser by calculating their standardized score z_i, and (2) normalizing the scores of the ICGs by introducing a factor that depends on the weights assigned to different descriptive attributes. The second approach takes advantage of specific information about ICGs. Such information is not available about parse trees, which motivates the use of the more generic first approach.

Adjusting parse tree probabilities. Given a probability p_i returned by the parser, we calculate its z-score $z_i = (p_i - \mu)/\sigma$, where μ is the mean and σ the standard deviation of the probabilities returned by the parser for our development corpus. The z_i scores are then transformed to the (0, 1] range using a sigmoid function $z_i^{Norm} = 1/(1 + e^{-z_i})$, which yields scores that increase linearly between the values [− 2, 2] of z_i.

Normalizing ICG scores. The ICG scores obtained by the Multiplicative scheme are often in a small band in a very low range, while the ICG scores obtained by the Additive scheme are typically greater than 1. In order to expand the range of the former, and map the latter into the (0, 1] range, we incorporate the following normalizing factor φ into their formulation:

$$\varphi=\sum_{i=1}^{M}\sum_{j=1}^{M}W_{\text{loc}}\delta(\text{loc}(u_i,u_j))+\sum_{i=1}^{M}\{W_{\text{lex}}+W_{\text{col}}\delta(u_i,colour)+W_{\text{size}}\delta(u_i,size)+W_{\text{unk}}\delta(u_i,unk)\}$$

This factor is incorporated into the Multiplicative and Additive schemes as follows:

- **Multiplicative scheme.**

$$\text{Sc}_{\text{MULT}}(I|U,\mathcal{C})=\text{Sc}_{\text{MULT}}(I|U,\mathcal{C})^{1/\varphi} \tag{7}$$

- **Additive scheme.**

$$\text{Sc}_{\text{ADD}}(I|U,\mathcal{C})=\frac{1}{\varphi}\text{Sc}_{\text{ADD}}(I|U,\mathcal{C}) \tag{8}$$

4.1.3. *Order of adjustments and normalization*

The above adjustments are made in the following order: the parse tree probabilities are adjusted before using them in the computation of the scores of UCGs, which takes into account multiple parents. The normalization factor φ is applied to adjust the score of an ICG derived from a single UCG, and the final score of this ICG is derived by considering multiple parents.

4.2. *Score estimation using distance-based semantics*

Modern approaches advocate the estimation of scores such as those in Eqs. (3) and (4) from data. However, it may be quite difficult to perform data-driven estimation of the similarity between the lexical item, colour, size and location in a description and the corresponding attributes of candidate objects in the physical world, and it may also be unnecessary, as the functions for performing this estimation have generally accepted distance-based semantics. For instance, a referent is deemed to be *on* a landmark when the bottom of the referent rests on the top of the landmark. As another example, given a request for a blue mug, a royal blue mug should have a higher match score than a green mug. However, as noted by Coventry and Garrod (2004), the details of these perceptions may vary between people, e.g., while it is generally agreed that two adjacent objects are *near* each other, the perception of when they stop being near depends on the individual and the context. Still, there is general agreement about relative attribute values, e.g., which object is nearer.

On the basis of these observations, we devise simple distance-based comparison functions to estimate the extent to which candidate objects match a description. These functions are largely based on approximations of the physical properties of objects, and implement generally accepted distance-based semantics. Our approach is similar to that adopted by Funakoshi et al. (2012), while significantly expanding the range of functions being handled. Table 2 displays an overview of our intrinsic and positional functions. The intrinsic functions, which estimate the scores in lines 2 and 3 of Eqs. (3) and (4), are described in detail in Appendix A. Our functions for positional relations, which resemble those by Coventry and Garrod (2004) and Kelleher and Costello (2008), estimate the scores in line 1 of Eqs. (3) and (4), and are detailed in Appendix B. As seen in Section 5, despite being rather coarse grained, these functions produce reasonable interpretations for the descriptions in our corpus.

5. Evaluation

Given a description, ideally a *Spoken Language Understanding (SLU)* system should understand what a person understands. We devised two experiments to assess *Scusi?*'s performance: *Interpretive* and *Generative*, where we compared complete interpretations (ICGs) generated by *Scusi?* to interpretations preferred by human addressees. These experiments in combination pose more stringent requirements than simply finding the correct referent (without identifying the correct landmark or positional relation) or employing IR-based metrics that compare the components of an interpretation generated by an SLU system with those of a reference interpretation (DeVault et al., 2009; Hirschman, 1998; Jokinen and McTear, 2010; Black et al., 2011).

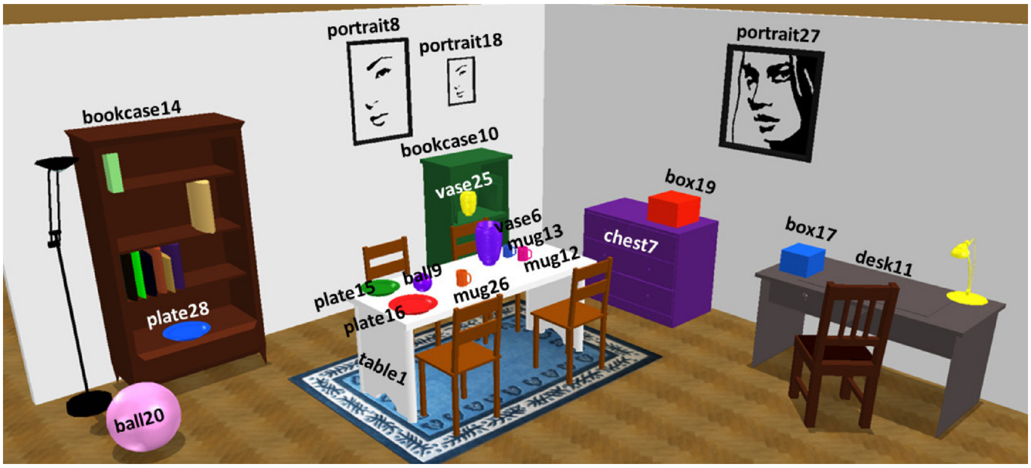


Fig. 2. Room with labeled objects for the Interpretive experiment. (For interpretation of the references to colour in the text, the reader is referred to the web version of the article.)

- **Interpretive** experiment – the participants and *Scusi?* were the addressees, and were given written descriptions generated by the authors. *Scusi?*'s confidence in its interpretations was then compared with the preferences of the participants (Section 5.3).
- **Generative** experiment – the participants were the speakers, generating free-form spoken descriptions, and *Scusi?* and expert annotators (the authors) were the addressees. *Scusi?*'s performance was evaluated on the basis of the rank of the correct interpretations it generated (Section 5.4).

In the Interpretive experiment, *Scusi?*'s performance was tested on textual input, while the Generative experiment was conducted both on input obtained from the ASR and on text (perfect ASR). *Scusi?* was set to generate at most 300 ICGs and sub-interpretations (texts, parse trees and UCGs) in total for both text and speech for each description, and its buffer capacity was set to $k_{\max} = 400$ for each sub-interpretation (Algorithm 1, Section 3).[3]

5.1. Description accuracy and Scusi?'s knowledge

People's understanding and that of an SLU system are affected by the precision of a description, i.e., ambiguous descriptions yield more than one suitable interpretation, and inaccurate descriptions yield interpretations that do not match the description or the real world perfectly. In addition, *Scusi?*'s vocabulary coverage and grammatical competence is lower than that of people, which also affects its performance. Hence, for our evaluation, we have found it useful to distinguish between two aspects of a description: *accuracy* and *knowledge*.

- **Accuracy** indicates whether a description represents an intended object precisely and unambiguously. For instance, when intending a blue plate, "the blue plate" is a precise and unambiguous description if there is only one such plate in the room, it is an ambiguous description if there is more than one such plate, and it is imprecise if all the plates in the room are of another colour. We distinguish between *perfect* and *imperfect* descriptions in terms of their accuracy. A description is *perfect* if it matches at least one object in the current context in every respect. In this case, an SLU module should produce one or more interpretations that match the description perfectly. If every object in the context mismatches a description at least in one aspect, the description is *imperfect*. In this case, we consider *reasonable* interpretations (that match the description well but not perfectly) to be the Gold standard. The *number* of Gold interpretations is an attribute of accuracy: a description may match (perfectly or imperfectly) 0, 1 or more than 1 interpretation, e.g., "the mug on the table" has three correct interpretations (*perfect* > 1) in the context of Fig. 2, while "the pink ball on the table" has two imperfect interpretations (*imperfect* > 1): `ball20`–*Location_on*–`table1`

[3] It is worth noting that *Scusi?*'s performance is quite insensitive to buffer size $200 \le k_{\max} \le 400$.

Table 3
Weights learned by steepest ascent hill climbing.

Combination scheme	Description type	Output of interpretation stages				Descriptive attributes				
		Text W_T	Parse tree W_P	UCG W_U	ICG W_I	Lexicon W_{lex}	Colour W_{col}	Size W_{size}	Position W_{loc}	Unknown W_{unk}
Multiplicative	Text	–	1	1	2	2	2	2	1	1
	ASR	2	1	3	1					
Additive	Text	–	4	1	4	2	1	2	1	1
	ASR	2	1	1	4					

and `ball19`–*Location_on*–`table1`. The first interpretation matches the UCG perfectly, but not the context, as the ball is not on the table; while the opposite happens for the second interpretation, owing to the mismatch between the requested pink colour and the purple colour of the ball that is on the table. If the pink ball (`ball20`) was adjacent to the table, a third interpretation `ball20`–*Location_near*–`table1` would also be a plausible imperfect match, where the imperfection stems from the mismatch between the *on* requirement in the UCG and the positional relation *Location_near*, which matches the context. In our Generative experiment, ambiguous (>1) descriptions constitute 17.8% of our corpus, and imperfect descriptions 16.5% (Section 5.4).

- **Knowledge** indicates how much an SLU module knows about different factors that affect the interpretation process, e.g., vocabulary, grammar or positional relations. A description is *known* to *Scusi?* if all the lexical items in the description appear in its vocabulary, the parser can generate correct parse trees from the description, *Scusi?* can translate the grammatical constructs in a parse tree to a UCG, and the prepositional phrases pertaining to positional relations in the description are understandable in terms of *Scusi?*'s geometric knowledge. Otherwise, the description is considered *unknown*. This category, which constitutes 27.2% of our corpus, is described in further detail in Section 5.4.2.

5.2. *Weights for interpretation stages and descriptive attributes*

As mentioned in Section 4, the effect of *Scusi?*'s interpretation stages and of the descriptive attributes in a referring expression on the score of an interpretation is represented by means of weights (Eqs. (1), (3) and (4)). If we had a single differentiable function for all descriptions that represents the goodness of each interpretation, we could perform optimization in a continuous space to learn these weights. However, it is not possible to define such a function, as the merit of an interpretation depends on the description from which it was generated. Still, one can apply search techniques to find a set of weights that optimizes performance over an entire dataset. To this effect, we conducted a modest experiment on a development corpus of 62 descriptions, where we employed two irrevocable search algorithms to learn these weights, viz a genetic algorithm and steepest ascent hill climbing; the fitness function was average *Normalized Discounted Cumulative Gain* @10 (*NDCG@10*) (Järvelin and Kekäläinen, 2002) (Section 5.4.3). To reduce the training time, we first learned the weights of under the Text condition (mishearing colour and size due to ASR error would turn them into other types of words, thus affecting the weight learning process), and then used the results of this experiment to learn the weights of the interpretation stages.

The best results were obtained with weights learned by steepest ascent hill climbing, where a weight was iteratively increased by 1 while keeping the other weights at their previous value. This process continued until performance did not improve, at which point the next weight was considered. We denote the version of *Scusi?* that uses these weights *WeightedScusi?*, and employ this version in the Interpretive and Generative experiments (Sections 5.3 and 5.4 respectively). These weights are displayed in Table 3 for the ASR condition and the Text condition under the Multiplicative and Additive attribute-combination schemes (Section 4).[4] As seen in Table 3, lexical item and size have a higher weight than position, and all the weights except colour are the same for both attribute-combination schemes.

[4] We also conducted experiments on a few small corpora where we considered increments of 0.5 for the weights, and several values for two configurable parameters: $k_{\max}$ and number of iterations. We found that the number of iterations affects performance, and that the learned weights are sensitive to the training data, which indicates that a larger training corpus is required to obtain stable weights.

Table 4
Descriptions for the Interpretive experiment with their accuracy.

#	Description	Accuracy
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

In contrast, the weights of the interpretation stages vary between these schemes. The textual and ASR inputs yield different UCG and ICG weights under the Multiplicative scheme, while they produce different parser weights under the Additive scheme.

Even though the Multiplicative attribute-combination scheme generally outperformed the Additive scheme on the development set, we employ both schemes in our experiments. In the Interpretive experiment, where the input is textual (Section 5.3) and in the textual Generative experiment (Section 5.4), *Scusi?* was run with the Text weights, and in the ASR-based Generative experiment, *Scusi?* was run with the ASR weights (Section 5.4).

5.3. *Interpretive experiment*

This experiment tests whether *Scusi?*'s understanding of descriptions that are mainly imperfect or ambiguous matches the understanding of a relatively large population.

5.3.1. *Experimental setup*

The trial consists of a Web-based survey where participants were shown a picture of a room comprising 29 objects of which 20 were labeled (Fig. 2), and given nine *known* written descriptions generated by the authors. The trial focuses on descriptions that are imperfect or ambiguous, as they pose a greater challenge to people than perfect descriptions: four descriptions are *imperfect* > 1, two are *perfect* > 1, one is *imperfect* = 1, and two are *perfect* = 1 (Column 2 in Table 4 shows the descriptions and Column 3 indicates their accuracy category). The imperfect descriptions play off different attributes against each other as follows: Description #2 plays off colour against size, Descriptions #4 and 5 play off colour against location, Description #7 plays off colour and location against size, and Description #8 plays off size against location. While these descriptions may seem contrived, they could plausibly be generated by a user who is not in the room, and misremembers the attributes or placement of objects. In addition, as mentioned in Section 5.1, imperfect or ambiguous descriptions constitute over 30% of our corpus (note that some descriptions are both imperfect and ambiguous).

For each description, participants were instructed to give a score between 0 and 10 to each of the labeled objects in the room, where 10 corresponds to a perfect match and 0 means that an object is irrelevant to the description – participants were told that entries left blank would receive a score of 0. The participants were asked to assign a score to each object, rather than only to their preferred object(s), as this forces them to consider partial matches between an object and a description, which in turn sheds light on the appropriateness of *Scusi?*'s scores and rankings.

5.3.2. *Data collection and analysis*

46 people participated in our study, resulting in 46 × 20 scores for each description. We calculated the mean and the median of the scores for each object across participants. There was some variation among the scores assigned by the participants to their preferred interpretations (standard deviation of both mean and median was less than 2 for six descriptions, and less than 3 for all descriptions), which may be attributed to some disregard for the instructions (e.g., some participants thought that a good score was 5, and some selected the landmark as well as the referent).

Table 5
Results of the Interpretive experiment.

#	Survey	*WeightedScusi?* I_1	*WeightedScusi?* I_2	*WeightedScusi?* I_3
1.	plate16	**plate16**/plate15 ***Loc_near*** **ball9**	plate15/**plate16** ***Loc_near*** **ball9**	plate28 *Loc_near* ball20
2.	box17	**box17**	box19	plate28, carpet23, mug13
3.	plate16	mug26	**plate16**	mug12
4.	bookcase14	bookcase10 *Loc_under* portrait18	bookcase10 *Loc_under* portrait8	**bookcase14** *Loc_under* portrait27/ portrait18/portrait8
5.	mug26	**mug26** ***Loc_near*** **vase6**	mug12 *Loc_near* vase6	mug13 *Loc_near* vase6
6.	plate28	**plate28**/plate16	plate16/**plate28**	plate15
7.	bookcase10	bookcase14 *Loc_near* chest7	**bookcase10** ***Loc_near*** **chest7**	bookcase14 *Loc_near* lamp22
8.	ball9	**ball9** ***Loc_on*** **table1**	ball20 *Loc_on* desk11/table1	ball20 *Loc_on* table1/desk11
9.	portrait18	**portrait18** ***Loc_above*** **bookcase10**	portrait8/portrait18 *Loc_above* bookcase10	portrait27 *Loc_above* bookcase10/ bookcase14 / portrait8 *Loc_above* bookcase14

The results of our survey for *imperfect* descriptions indicate that colour takes precedence over size (Description #2 in Table 4), thus confirming Gatt et al.'s (2007) observations. In addition, our participants prioritized colour over location (Descriptions #4 and #5), but gave location a higher priority than size (Description #8).

5.3.3. Evaluation metrics

Scusi?'s understanding of each description was compared with that of our trial subjects by calculating the Pearson correlation coefficient and the Spearman rank correlation coefficient between the average (and median) of the scores of the subjects' ratings for each labeled object $O_1, \ldots, O_{20}$, and the score assigned by *Scusi?* to the highest-ranking ICG with a head object that matches $O_1, \ldots, O_{20}$ respectively. For instance, given Description #1 "the plate next to the ball", under the Multiplicative attribute-combination scheme, *Scusi?* ranked `plate15-near-ball9` and `plate16-near-ball9` equal first with a score of 4.04×10^{-8} (the plates are equidistant from the ball), and `plate28-near-ball20` second with a score of 3.71×10^{-8}, followed by ICGs comprising non-plate objects that are near balls, e.g., `mug26` and `bookcase14`, and so on.[5] We computed the Pearson correlation and the Spearman rank correlation between these scores and the average and median scores assigned by our participants to `plate16` (average 7.2), `plate15` (7.1), `plate28` (6.6), `mug26` (0.7), `bookcase14` (0.5), etc.

5.3.4. Results

Table 5 compares the results of our survey with the results obtained by *WeightedScusi?* under the Multiplicative attribute-combination scheme. Column 2 displays the object preferred by the trial subjects, and Columns 3–5 show the top three interpretations preferred by *WeightedScusi?* (I_1–I_3). Matches between the system's output and the averaged participants' ratings are boldfaced. As seen in Table 5, *WeightedScusi?*'s top-ranked interpretation matches our participants' preferences in six cases, of which two have the same score as the system's second-ranked interpretation (Descriptions #1 and 6). *WeightedScusi?*'s second-ranked interpretation matches the participant's preferences in two cases (Descriptions #3 and 7).

Table 6 shows the Pearson correlation coefficient and the Spearman rank correlation coefficient between *WeightedScusi?*'s scores and the mean and median of the participants' scores for each labeled object, compared to the correlations obtained by two baselines: *UnityScusi?*, where all the weights equal 1; and *Top1-Equal-Score* – a baseline that maintains only top same-score interpretations, which is a lenient version of a top-1 baseline. As seen in Table 6, both versions of *Scusi?* yielded a similar performance, and significantly outperformed the *Top1-Equal-Score* baseline. In most cases, except for Spearman rank correlation coefficient for *WeightedScusi?*, the Additive attribute-combination scheme performed at least as well as the Multiplicative scheme. Specifically, the top-ranked interpretations produced

[5] The scores produced by *Scusi?* are quite low, but since the ranking of the interpretations is relative, there is no need to take further action at this stage. However, as shown by Lee (2014), an accurate indication of the quality of an interpretation is essential for dialogue-state tracking in a Spoken Dialogue System. In the future, we will re-calibrate the scores of the interpretations by dividing them by the score of a hypothetical perfect interpretation, which will indicate how far *Scusi?*'s interpretations are from an ideal one.

Table 6
Performance for the Interpretive trial: *WeightedScusi?*, *UnityScusi?* and *Top1-Equal-Score*.

Attribute combination scheme	*WeightedScusi?*				*UnityScusi?*				*Top1-Equal-Score*			
	Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Multiplicative	0.87	0.84	0.70	0.78	0.85	0.83	0.68	0.72	0.64	0.63	0.44	0.58
Additive	0.89	0.87	0.69	0.75	0.87	0.83	0.71	0.72	0.72	0.73	0.45	0.59

by *WeightedScusi?* under the Additive scheme matched the participants' preferences in eight cases, compared to the six cases matched by the top-ranked interpretations generated under the Multiplicative scheme (Table 5). This may be attributed to the fact that the Additive scheme is more forgiving of inaccuracies than the Multiplicative scheme, and five of the nine descriptions are inaccurate. The lower Spearman rank correlations (compared with the Pearson correlations) are mainly due to the fact that some participants gave non-zero scores to implausible interpretations (e.g., they selected the landmark of a description as well as the referent), while both versions of *Scusi?* produce many interpretations with the same low rank (implausible interpretations get a score close to 0), and the *Top1-Equal-Score* baseline returns at most a few interpretations (with the remaining options receiving a score of 0).

The main discrepancies between *WeightedScusi?*'s preferred interpretations and our participants' preferences are due to the following reasons:

- Description #3 “the red dish” – according to Leacock and Chodorow's similarity metric (Appendix A), a mug is more similar to a dish than a dinner plate, and according to the CIE94 colour scheme, orange is sufficiently similar to red for the mug to be the winning candidate. However, our trial subjects thought that only plates can be dishes.
- Description #4 “the brown bookcase under the portrait” – `bookcase14` is ranked ahead of `bookcase10` by the Additive scheme, while the Multiplicative scheme penalizes heavily attributes that do not match reality (Section 4), thereby significantly reducing the score of `bookcase14`, which is not under any portrait. In contrast, `bookcase10` is under a portrait, and its shade of green is considered somewhat similar to the requested colour brown.
- Description #6 “the large plate” – our participants perceived `plate28` to be larger than `plate16` although they are the same size, and hence have the same score.
- Description #7 “the large green bookcase near the chest” – the Multiplicative scheme ranks `bookcase14` ahead of `bookcase10`, because `bookcase14` matches all the requested attributes to some extent, while `bookcase10` is unlikely to be considered large, and hence has a low size-match score. Here too the Additive scheme is more forgiving, ranking `bookcase10` first.

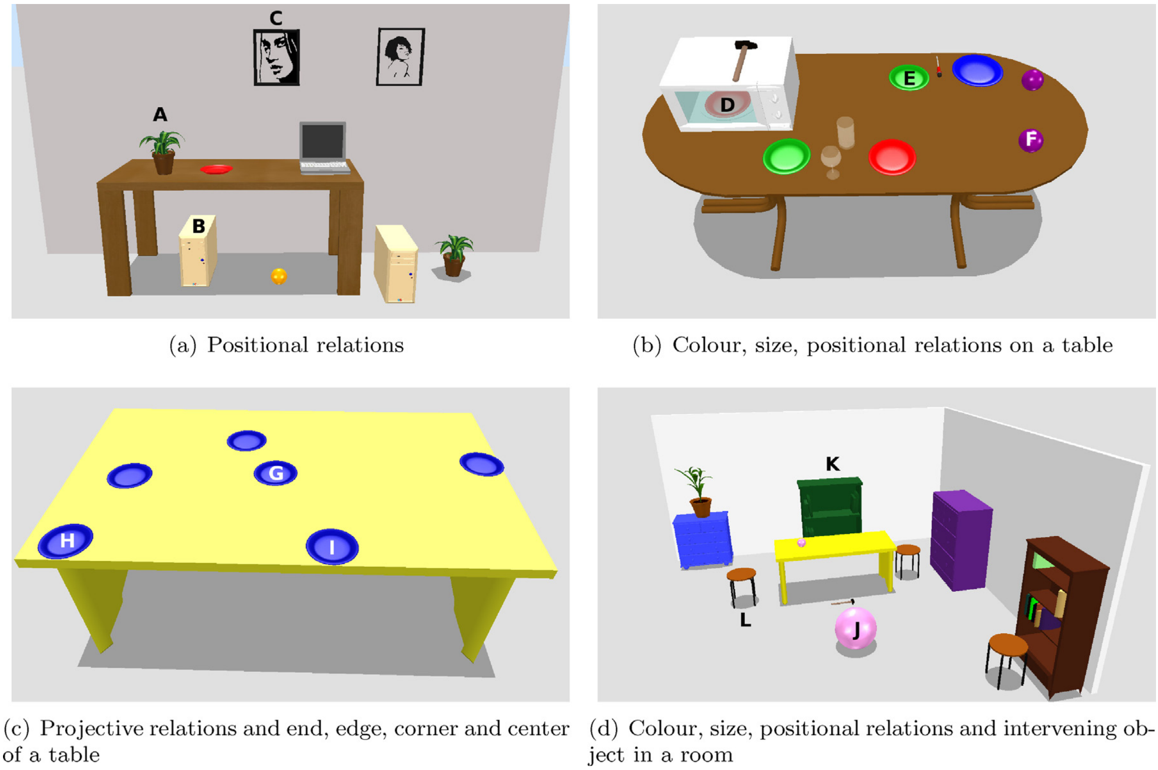
Thus, according to this trial, which focuses on imperfect and ambiguous descriptions, both versions of *Scusi?* satisfy our original requirement for reasonable behaviour (Section 1), with the Additive scheme slightly outperforming the Multiplicative scheme.

5.4. Generative experiment

This experiment assesses *Scusi?*'s performance when given free-form descriptions, both spoken and textual (perfect ASR). As for the Interpretive experiment, we compare the performance obtained by *WeightedScusi?* with that obtained by our two baselines, *UnityScusi?* and *Top1-Equal-Score*, under the Multiplicative and the Additive attribute-combination schemes.

5.4.1. Experimental setup

In this experiment, trial subjects generated free-form, spoken descriptions to identify 12 designated objects (labeled A to L) in four scenarios – three objects per scenario, each of which contains between 8 and 16 objects including the speaker (Fig. 3). These scenarios, which were designed to test different functionalities of our system, contain various distractors in terms of colour, size and position, and an intervening object between a referent and a potential landmark. Specifically, Fig. 3(a) contains two pot plants, three computers and two portraits, requiring positional descriptions to



(a) Positional relations

(b) Colour, size, positional relations on a table

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

(d) Colour, size, positional relations and intervening object in a room

Fig. 3. Scenarios for the Generative experiment. (For interpretation of the references to colour in the text, the reader is referred to the web version of the article.)

distinguish between a labeled referent and a distractor; Fig. 3(b) requires different positional relations, possibly in combination size or colour; Fig. 3(c) was designed to test projective and topological relations, as all the plates are equal; and Fig. 3(d) shows similar objects that differ with respect to some attributes in a realistic setting, as well as an intervening object (the hammer) between referent J and the yellow table, which is a potential landmark. These scenarios significantly extend the range of descriptions considered to date, e.g., (Zukerman et al., 2008; Moratz and Tenbrink, 2006; Funakoshi et al., 2012; Kelleher and Costello, 2008).

The annotators provided the Gold standard interpretations for each description on the basis of what they understood (rather than by using the designated referents), since, as indicated in Section 5.1, people sometimes generated ambiguous or imperfect descriptions. Each annotator handled half of the descriptions, and the other annotator verified the annotations. Disagreements were resolved by consensus.

5.4.2. *Data collection and analysis*

Our study had 26 participants (half were native English speakers), who were academic and administrative staff and graduate students in the Faculty of Information Technology at Monash University.

At the beginning of the experiment, we explained the task to our trial subjects (to describe labeled objects in forthcoming pictures), and informed them about the understanding capabilities of the system: *Scusi?* understands names of things, colours and sizes, and positional relations with landmarks, but it does not understand the function (e.g., "*coffee* mug"), texture (e.g., "*shiny* vase"), composition (e.g., "*ceramic* cup") or shape (e.g., "*round* dish") of objects, or positional relations without landmarks (e.g., "the ball *in the middle*"). We also asked the participants to utter their descriptions without hesitations or pauses, as the ASR does not cope with disfluencies, and informed them that they can repeat or modify a description at most twice (maximum three descriptions in total for one referent).

The experiment was conducted in a quiet room, where only one participant and the experimenter were present. The participants employed a hand-held microphone when interacting with the ASR, and spent about 10 min training the ASR prior to the actual recording of their descriptions. We then showed them the pictures in Fig. 3, and asked them to designate each of the referents labeled A to L by means of a description (the order of presentation of the pictures was rotated, yielding four test groups). During their interaction with the ASR, the participants could view the candidate texts produced by the ASR for each spoken description, which informed their decision to repeat or modify a description. This interaction was separate from the activation of *Scusi?*. That is, for each spoken description, the ASR produced a file containing up to 50 alternative texts. These files were submitted to *Scusi?* for processing in batch mode. In addition, in order to determine the influence of ASR error on *Scusi?*'s performance, we also submitted files containing correct textual transcriptions of the participants' spoken descriptions.

The participants generated a total of 432 spoken descriptions (average length was 7.8 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 analyzed the remaining 400 descriptions on the basis of *Scusi?*'s knowledge.[6] 290 of these descriptions were *known*; the remaining 110 *unknown* descriptions were categorized into four classes, of which only the first one is representable by *Scusi?*: *out of vocabulary words or out of scope position modifiers*, *complex positional relations*, *positional relations without a landmark* and *out of grammar relations*.

- *Out of vocabulary (OOV) words or out of scope (OOS) position modifiers* – these are descriptions for which, despite some words not being understood by *Scusi?*, in principle *Scusi?* could build a Gold standard interpretation that constitutes a *reasonable* representation of the referent (although there is no guarantee that it will do so). Examples are: "the *motherboard* underneath the table" to designate object B in Fig. 3(a), and "the blue plate in the *left near* corner of the table" to refer to object H in Fig. 3(c) (both "left" and "near" are known by *Scusi?*, but not as modifiers of a position). Even though *Scusi?* does not know what a motherboard is, it could postulate two candidate same-score interpretations comprising the two objects under the table in Fig. 3(a), of which only one is intended. Similarly, *Scusi?* can identify the plate in a corner of the yellow table in Fig. 3(c), even though it does not know what "left near" is, producing the interpretation `plate3`–*Location_corner*–`table1`, which is the intended one (if there had been plates in other corners, *Scusi?* would have produced additional interpretations that have the same score as the intended one). Our evaluation metrics take such situations into account by assigning fractional scores to correct interpretations that have the same score as wrong interpretations (Section 5.4.3). 12.8% of the descriptions in our corpus belong to this category.
- *Complex positional relations*, where the complexity of a positional relation is beyond *Scusi?*'s geometric knowledge or grammatical ability, e.g., "the blue plate *on the right of the center of* the table" and "the green bookcase *at the far side* of the room". This category comprises 7.5% of the descriptions in our corpus.
- *Positional relations without a landmark*, such as "the ball *in the center*". 5.2% of the descriptions in our corpus are of this type.
- *Other out of grammar (OOG) relations*, such as prepositional phrases starting with "of" or "with" and subordinate clauses, e.g., "the picture *of the face*", "the plant *with the leaves*" and "the pink ball *that is* on the floor". 5.3% of the descriptions in our corpus belong to this category.

Two main approaches for handling descriptions in *unknown* categories are: improving the competencies of an SLU system, e.g., expanding *Scusi?*'s vocabulary, extending its geometric knowledge, or improving the algorithm that translates parse-trees to UCGs; or circumventing the problem, i.e., judiciously disregarding the parts of an utterance that are not understood, and generating clarification questions as necessary. Analysis of the frequency of these categories enables us to determine when the first approach is warranted. However, even though one can always improve the capabilities of an SLU system, it is likely that people will sometimes use expressions or terms that are not known by the system. In this research, we have adopted a limited version of the second approach, whereby *Scusi?* generates interpretations for the 51 descriptions in the OOV/OOS *unknown* category, which can be represented within *Scusi?*'s framework. This enables us to make an initial assessment of how well *Scusi?* copes with such unknown descriptions. In the future, we propose to extend this approach to handle missing landmarks, complex positional relations and OOG relations. However, for the time being, the 59 descriptions in these *unknown* categories, which cannot be represented within *Scusi?*'s framework, are removed from our corpus. Table 7 shows the breakdown of these descriptions, which

[6] The dataset comprising these descriptions and the corresponding knowledge base and images is available upon request from the authors.

Table 7
Unknown descriptions that *Scusi?* cannot represent.

	Positional relation		Out of Grammar	
	Complex	No landmark	"with"/"of"	Other
perfect = 1	10	0	6	6
perfect > 1	3	0	4	0
imperfect = 1	15	20	1	4
imperfect > 1	2	1	0	0
Total	30	21	11	10

Table 8
Breakdown of descriptions in terms of accuracy and knowledge.

	Known	Unknown
perfect = 1	214	39
perfect > 1	47	6
imperfect = 1	22	5
imperfect > 1	7	1
Total	290	51

Table 9
Categorization of ASR errors.

Error type	Example	Percentage
Wrong determiner	"the plate" → "*that* plate"	7.5
Same PoS for object, landmark or position	"the ball" → "the *bowl*"	55.0
Different PoS for object, landmark or position	"the blue plate" → "*to build played*"	28.9
Additional words (i.e., noise)	"*and* the picture on the wall"	8.6

constitute 14.8% of our initial corpus (13 *unknown* descriptions were assigned to more than one heading, e.g., "the plate *in the far left in front*" contains a complex positional expression with no landmark).

Table 8 displays the frequencies of the four accuracy classes (*perfect* = 1 and >1 and *imperfect* = 1 and >1) and two knowledge classes (*known* and *unknown*) for the remaining 341 descriptions (Section 5.1). For instance, the top row shows the *perfect* = 1 descriptions, of which 214 are *known* and 39 are *unknown*. The low frequencies of the *imperfect* > 1 classes suggest that people rarely generate descriptions that are both ambiguous and inaccurate.

ASR performance. The average *Word Error Rate* (*WER*) was 30.4% for the top-ranked ASR output and 31.8% for the top three alternatives returned by the ASR for the 341 descriptions that have Gold standard interpretations, and marginally higher for the 400 descriptions in the entire corpus. The ASR produced the correct interpretation at the top rank for 13.8% of the 341 descriptions, in the top three ranks for 26.1% of these descriptions, and at any rank for 35.8%. These percentages are slightly lower for the 400 descriptions in the entire corpus. Focusing on the 252 descriptions that had no correct ASR output in the top three ranks, we categorized the ASR errors into four classes as shown in Table 9 (59 of these descriptions had more than one ASR error). This categorization was performed over the most promising ASR outputs. As seen in Table 9, about 63% of the errors made by the ASR (the first two categories) retained the correct Part-of-Speech (PoS) for a misheard word, which generally leads to a correct parse tree (modulo the erroneous word(s)), and improves the chances of generating a reasonable ICG. This is not the case for the remaining 37% of the errors (the last two categories), which result in an incorrect parse tree. These hypotheses are validated in Section 5.4.4.

5.4.3. Evaluation metrics

We employed two metrics to evaluate our system: *FRecall@K* and *NDCG@K*, where the @*K* component evaluates performance for different cut-off ranks *K*. When $K = \infty$ all the interpretations returned by *Scusi?* are considered.[7]

- ***Fractional recall* (*FRecall*) @*K*** – This measure is similar to the conventional *Recall* metric, which is defined as follows:

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

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

However, *FRecall* takes into account the fact that an N-best SLU module may return several interpretations with the same score, 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 same-score interpretations. *FRecall@K* allows us to represent the arbitrariness of the ranked order of same-score interpretations as follows:

$$FRecall@K(d) = \frac{\sum_{j=1}^{K} fc(I_j)}{|C(d)|}, \tag{9}$$

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

$$fc(I_j) = \frac{c_j}{h_j - l_j + 1}, \tag{10}$$

where l_j is the lowest rank of all the interpretations with the same score 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 *FRecall@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 $fc(I_j)$ as a measure of the relevance of interpretation I_j, we obtain

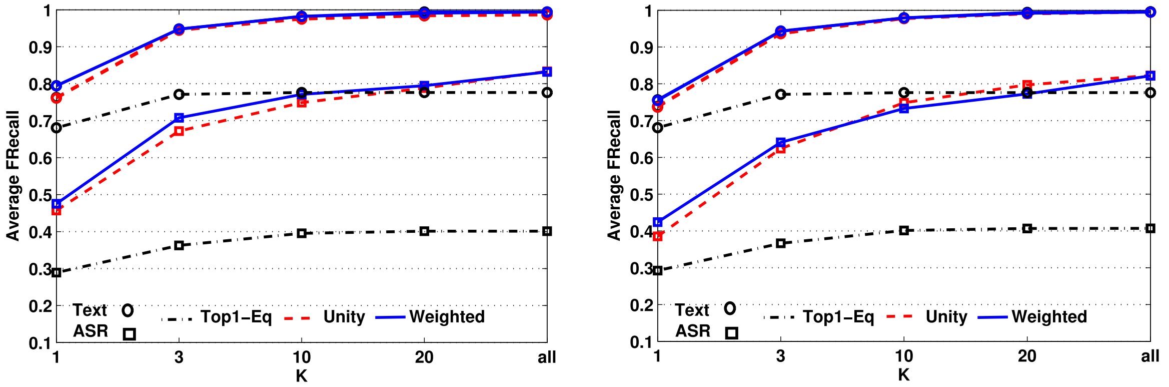
$$DCG@K(d) = fc(I_1) + \sum_{j=2}^{K} \frac{fc(I_j)}{\log_2 j}. \tag{11}$$

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

$$NDCG@K(d) = \frac{DCG@K(d)}{1 + \sum_{j=2}^{\min\{|C(d)|, K\}} \frac{1}{\log_2 j}}. \tag{12}$$

Note that *FRecall@K* is computed in relation to the total number of correct interpretations, while *NDCG@K* considers the minimum of *K* and this number (Eqs. (9) and (12) respectively). Also, we do not use the Information Retrieval (IR) precision measure because, contrary to IR settings, where typically there are many documents relevant to a query, in language understanding situations, there is often only one correct interpretation for an utterance (Table 8). Hence, precision does not reflect a system's performance.

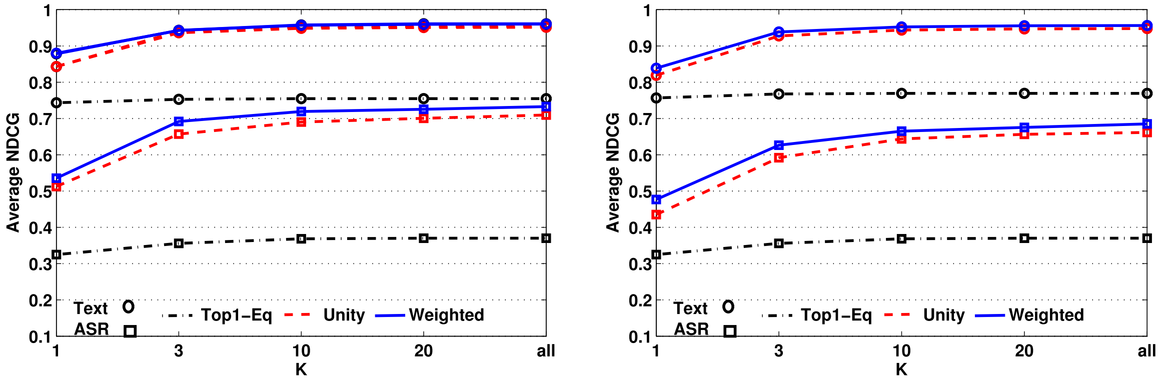
[7] Recall that the *Top1-Equal-Score* baseline often returns one interpretation. Therefore, when evaluating this baseline, we use the same value of *K* as that employed for *Scusi?*.



(a) Multiplicative scheme: *FRecall@K*

(b) Additive scheme: *FRecall@K*

Fig. 4. Performance for the Generative trial: *FRecall@K* for *WeightedScusi?*, *UnityScusi?* and *Top1-Equal-Score*.



(a) Multiplicative scheme: *NDCG@K*

(b) Additive scheme: *NDCG@K*

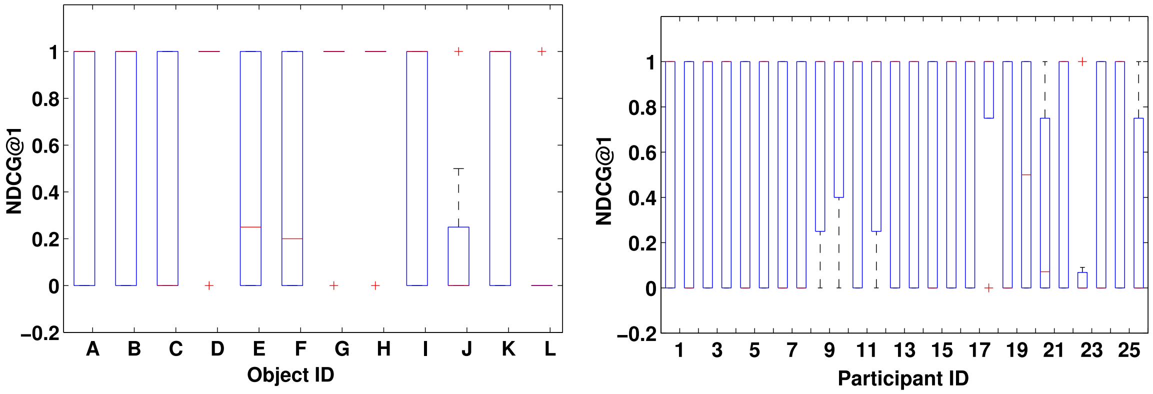
Fig. 5. Performance for the Generative trial: *NDCG@K* for *WeightedScusi?*, *UnityScusi?* and *Top1-Equal-Score*.

5.4.4. Results

As mentioned above, we compared the performance of *WeightedScusi?* with that of our baselines, *UnityScusi?* and *Top1-Equal-Score*. *WeightedScusi?* and *UnityScusi?* were compared under both the Multiplicative attribute-combination scheme and the Additive scheme for inputs obtained from the ASR and for textual inputs (perfect ASR). However, we employed the best version of the *Top1-Equal-Score* baseline for our comparisons. This version uses unity weights under the Additive attribute-combination scheme for input from the ASR and for textual input. Figs. 4 and 5 depict *WeightedScusi?*'s performance, compared with that of *UnityScusi?* and the *Top1-Equal-Score* baseline, in terms of average *FRecall@K* and average *NDCG@K* respectively for the 341 descriptions with Gold standard interpretations. Statistical significance was calculated using the two-tailed Wilcoxon signed rank test. We now discuss several aspects of *Scusi?*'s performance.

Scusi? versus the Top1-Equal-Score baseline. As seen in Figs. 4 and 5, both versions of *Scusi?* significantly outperform this baseline for both textual input and input obtained from the ASR, under both the Multiplicative attribute-combination scheme and the Additive scheme, for all values of K (p-value $\ll 0.01$). The superior performance of *Scusi?*, even for $K = 1$, is due to its stochastic operation, which may expand options that initially appear inferior, allowing them to generate descendants that may overtake the descendants of top-ranked sub-interpretations. For instance, a lower-ranked ASR output, which would never be inspected by the *Top1-Equal-Score* baseline, may be the ancestor of the final top-ranked ICG.

Effect of attribute-combination scheme together with weights of interpretation stages and descriptive attributes. *Scusi?*'s performance under the Multiplicative attribute-combination scheme was better than its performance under the Additive scheme for input from the ASR and textual input for both *UnityScusi?* and *WeightedScusi?*. Specifically,



(a) Referent effect on $NDCG@1$

(b) Participant effect on $NDCG@1$

Fig. 6. Effect of referent and participant on *Scusi?*'s performance.

for ASR input, *FRecall@K* under the Multiplicative scheme exceeded *FRecall@K* under the Additive scheme by 5 percentage points for $K \leq 3$, while this happened for all values of K for *NDCG@K*. However, none of these differences was statistically significant at p-value < 0.05. In addition, the effect of the attribute-combination scheme on *Scusi?*'s performance on textual descriptions was smaller.

Comparing *WeightedScusi?* with *UnityScusi?*, the weights assigned to the interpretation stages and descriptive attributes also had little effect on *Scusi?*'s performance on textual input, with *WeightedScusi?* statistically significantly outperforming *UnityScusi?* only for $K = 1$ under the Multiplicative attribute-combination scheme. However, this is not the case for input obtained from the ASR, where *WeightedScusi?* outperformed *UnityScusi?* under the Multiplicative scheme (statistically significant with p-value < 0.05 for *NDCG@K* for $K \geq 3$, and for *FRecall@K* for $K = 3$, 10) and under the Additive scheme (statistically significant with p-value < 0.05 for *NDCG@K* for $K \geq 10$, and for *FRecall@K* for $K = 1$, 20).

Effect of input quality: textual input versus input from the ASR. As expected, *Scusi?*'s performance on textual input significantly exceeded its performance on input obtained from the ASR. Specifically, for textual input, *WeightedScusi?* under the Multiplicative attribute-combination scheme yielded $FRecall@1 = 0.79$ and $FRecall@3 = 0.95$, compared to $FRecall@1 = 0.48$ and $FRecall@3 = 0.7$ for input obtained from the ASR (Fig. 4(a)); similar results were observed for *NDCG@K* (Fig. 5(a)). These results strongly suggest that improvements in ASR accuracy will yield significant improvements in *Scusi?*'s performance. Our results also indicate that *Scusi?* produced Gold standard interpretations for a substantial portion of the 341 descriptions that had ASR errors, as only 13.8% of the top-ranked ASR outputs, 26.1% of the top three ASR outputs and 35.8% of the ASR outputs at any rank had no errors (Section 5.4.2). In particular, *Scusi?* was able to overcome ASR errors when misheard words had the same PoS as the corresponding correct words (e.g., "flour" instead of "flower"), and descriptions contained redundant information, which is often the case (Dale and Reiter, 1995; Levelt, 1989).

As mentioned in Section 5.4.2, participants were allowed to repeat or rephrase a description up to two times. Our results show that *Scusi?* yielded substantial performance improvements for participants who repeated/rephrased five or more of their descriptions once, while repeating or rephrasing a second time did not yield improvements.

Effect of referent and participant. Fig. 6(a) and (b) respectively depicts *NDCG@*1 for each referent and each participant. As seen in Fig. 6(a), seven referents had a median of 1 for *NDCG@*1, while three referents had a median of 0 (referent C in Fig. 3(a), and referents J and L in Fig. 3(d)), which means that *Scusi?* failed to generate a top-ranked Gold ICG for most of the descriptions of these referents. For referents J and L, this mainly happened due to ASR error combined with insufficient information in the remainder of a description to enable *Scusi?* to discriminate between the referent and potential distractors. For example, when "the pink ball in the middle of the room" was misheard as "the *pig bowl* in the middle of the room", any object near the middle of the room became a plausible candidate. Since the room in Fig. 3(d) has objects that are closer to the middle than the target pink ball, e.g., the hammer, the table and the small pink ball, the score of the referent falls below the score of these objects. Referent C exhibited additional problems, namely a higher WER coupled with seven out of 22 descriptions having more than one landmark. The latter caused the parser to generate several candidate parse trees, each with its own UCG and ICG expansions. As a result, the Gold ICG was generated at the top rank for only two descriptions (three descriptions in the top three ranks). This observation reinforces the idea of applying a learning regime similar to that used to learn the weights of the interpretation stages and descriptive attributes (Section 5.2) in order to learn the cut-off thresholds described in Section 2.

Fig. 6(b) shows that 15 participants had a median *NDCG*@1 of 1, compared to no participants with a median WER of 0 for the top-ranked ASR textual output, thus providing further evidence of *Scusi?*'s ability to overcome ASR errors. It is also worth noting that participant 23, who could not be understood by the ASR, was included in our analysis, which adversely affects the results shown in Figs. 4 and 5.

Effect of description accuracy and Scusi?'s knowledge. As expected, *Scusi?*'s performance on *known* textual input was better than its performance on *unknown* textual input across all accuracy classes. However, *Scusi?*'s knowledge state had no effect on its performance on input obtained from the ASR for accurate ($perfect = 1$) descriptions (there are not enough descriptions in the other *unknown* categories to draw meaningful conclusions). This may be attributed to ASR error, which changes known words into unknown words, thus reducing the differences between the *known* and *unknown* categories.

It is worth noting that the *FRecall*@*K* values are similar to the *NDCG*@*K* values for the unambiguous accuracy classes ($perfect = 1$ and $imperfect = 1$) and for $K \geq 3$ for the ambiguous classes ($perfect > 1$ and $imperfect > 1$). However, the *FRecall*@1 values are about half of the *NDCG*@1 values for the ambiguous classes. This is an artefact of the way these metrics are calculated: the denominator of *FRecall*@*K* contains the total number of Gold standard interpretations (Eq. (9)), while the denominator of *NDCG*@*K* is the minimum of the number of Gold interpretations and *K* (Eq. 12).

Performance analysis. Under the Multiplicative attribute-combination scheme, *WeightedScusi?* failed to find a Gold standard interpretation for 25 textual (perfect ASR) descriptions at $K = 1$ (22 *known*), and for 12 textual descriptions at $K \leq 3$ (9 *known*). Five of these 12 descriptions included the word "hammer", e.g., "the green plate next to the hammer" (Fig. 3(b)), which was erroneously parsed as a verb; three descriptions were imperfect, e.g., "the computer behind the table" (Fig. 3(a)), and three were ambiguous, e.g., "the purple ball further away from the plate" (Fig. 3(b)), yielding plausible alternatives that were ranked ahead of the Gold standard interpretation; and one description, "the blue plate on the edge of the table closest to me near the middle of the table", which has three prepositional phrases, yielded multiple parse trees that were expanded before the Gold interpretation could be generated. As indicated above, *WeightedScusi?* found Gold-standard interpretations at ranks 2 or 3 for 13 descriptions. The interpretations that preceded these descriptions were quite plausible. For instance, descriptions containing different wordings of "the stool in front of the blue chest of drawers" (Fig. 3(d)) yielded the stool in front of the purple chest slightly ahead of referent K. *Scusi?* ranked this stool first because it is more "in front" of a chest than the referent according to the viewer-centered frame of reference in light of the speaker's position, and purple is sufficiently similar to blue. This example illustrates the need to consider additional frames of reference for projective relations (Section 4).

WeightedScusi? did not find Gold standard interpretations in the top three ranks for 27% of the spoken descriptions, of which 79 were *known* by *Scusi?* and 13 were *unknown*. Another 18, 9 and 11 Gold interpretations were respectively found for $K = 10$, 20 and ∞. *Scusi?*'s failure to find a top-three ranked Gold ICG for 92 descriptions may be attributed to the different interpretation stages as follows: 87 to ASR errors, two to the parser (two of the descriptions with "hammer" which were correctly heard by the ASR), one to the UCG conversion process, and two to incorrect positional descriptions (the third of the above-mentioned imperfect descriptions was not heard correctly by the ASR). The 87 ASR errors that caused *Scusi?* to fail are distributed as follows among the four error categories in Table 9: 5 due to *Wrong determiner* (23.8% of the descriptions that had this error), 52 due to *Same PoS* (33.8% of the descriptions that had this error), 51 due to *Different PoS* (63.0%), and 11 due to *Additional words* (45.8%). These results validate our hypothesis from Section 5.4.2 whereby it is easier for *Scusi?* to recover from ASR errors that retain the PoS of a spoken utterance.

6. Related research

This paper is at the intersection of research on understanding referring expressions, human–robot interaction, models of spatial reasoning, and performance evaluation, and is part of an increasing trend which harnesses numerical formalisms to handle the uncertainty inherent in real-world problems (Funakoshi et al., 2012; Lison and Kruijff, 2009; Ross, 2010).

Understanding referring expressions. The use of different attributes in object descriptions has been studied both in psychology and in *Natural Language Generation* (*NLG*) (Krahmer and van Deemter, 2012), but there is little related research in *Natural Language Understanding* (*NLU*). Several studies have found that people tend to include in their descriptions attributes that do not add discriminative power, e.g., (Dale and Reiter, 1995; Levelt, 1989), which can be partly explained by the incremental nature of human language production and understanding (Pechmann, 1989; Kruijff et al., 2007). van Deemter (2006) and Mitchell et al. (2011) considered the generation of descriptions that employ gradable attributes, obtained from numerical data, focusing on size-related modifiers; and Gatt et al. (2007) compared the performance of several generation algorithms with respect to a combination of features, viz colour, position (restricted to placement in a grid), orientation and size. Their algorithm produced descriptions similar to those generated by people when the priority order of the attributes was $colour \succ orientation \succ size$. In contrast, Herrmann and Deutsch (1976) found that there is no universally applicable priority ordering of attributes, and Dale and Reiter (1995) suggested empirical investigations to determine the priority order of attributes for different domains. This suggestion has been adopted and extended in this paper by applying search techniques to learn weights that reflect the importance of object attributes and processing stages in the interpretation of object descriptions (Section 5.2), and by considering two schemes for combining descriptive attributes (Section 4).

Methods for interpreting referring expressions in dialogue systems have been examined in Funakoshi et al. (2012) and Pfleger et al. (2003). Pfleger et al. (2003) used modality fusion to combine hypotheses from different analyzers (linguistic, visual and gestural), choosing as the referent the first object satisfying a differentiation criterion. As a result, their system does not handle situations where more than one object satisfies this criterion. Funakoshi et al. (2012) offered a unified formalism based on Bayesian Networks which handles descriptions, anaphora and deixis, and incorporated *reference domains* (Salmon-Alt and Romary, 2000), viz sets of referents presupposed at each use of a referring expression, to disambiguate referring expressions. Similarly to our approach, they estimate probabilities by means of heuristics. However, their mechanism currently deals with descriptions regarding simple geometric shapes, and it does not handle sequenced positional relations, e.g., “the ball near the plate on the table”.

Human–robot interaction. Moratz and Tenbrink (2006) defined *acceptance areas* for projective relations that are similar to ours, and conducted experiments to determine the types of instructions people give to a robot under a primed and an unprimed condition. They found that people who have not been primed tend to use OOV and OOG expressions, and to give imprecise projective relations – a situation that improves after participants are notified of the robot’s capabilities. Both Hüwel and Wrede (2006) and Skubic et al. (2004) employed semantic grammars and a frame-based representation to encode instructions given to a robot. Hüwel and Wrede focused on two tasks, viz showing a robot around a house and instructing a robot to water several plants, while Skubic et al. modeled positional descriptions and multimodal directional commands, focusing on unoccupied regions near objects (e.g., “the left of the pillar”). Hüwel and Wrede argued that semantic grammars enable them to overcome some ASR errors and fill in missing information, and employed a fusion mechanism to combine the words understood by their system. However, semantic grammars restrict the range of utterances that can be understood by an SLU system (Knight et al., 2001). Like *Scusi?*, the spatial computations performed by both of these systems use idealizations, such as bounding boxes and center of gravity (Appendix B). However, owing to their propositional knowledge representation, neither of these systems handles ambiguous descriptions. In contrast, like *Scusi?*, Ross’s system (Ross, 2010) maintains multiple interpretations and assigns them scores based on their likelihood, which enables his system to handle ambiguity and ellipsis. However, his domain of implementation was a sequence of simple moves, rather than *Scusi?*’s relatively complex descriptions.

Kruijff et al. (2007) and Kruijff et al. (2010) argued that the time needed to process an utterance can be reduced by using a “packed” representation, whereby if multiple alternative interpretations share the same sub-content, their condensed representation contains that content only once. Such an idea has been implemented in *Scusi?* for complete interpretation structures (i.e., ICGs), but could be extended to collapse similar sub-interpretations (i.e., ASR outputs, parse trees and UCGs). Further, Kruijff et al. suggested processing partial interpretations immediately across all stages to allow the interpretation search tree to be pruned early. Although in principle we favour ranking ill-matching interpretations over removing them, and making global, rather than local, decisions, we have employed some pruning (via the cut-off threshold described in Section 2) to ignore unpromising interpretations.

Spatial reasoning. Kelleher and Costello (2008) proposed a model that focuses on topological and projective descriptions, taking into account the presence of intervening objects and the salience of landmarks. During discourse interpretation, their system first selects objects that match type and colour specifications, and then applies their positional model to determine the intended referent. In contrast, our probabilistic formulation weighs the contribution of various factors to the goodness of an interpretation. In addition, unlike *Scusi?*, they considered restricted, artificial settings with only a few objects, and did not process sequenced positional relations.

Performance evaluation. SLU systems have typically been evaluated using IR-based measures where the semantic components in an interpretation returned by these systems are compared with those of a reference interpretation (DeVault et al., 2009; Hirschman, 1998; Jokinen and McTear, 2010; Black et al., 2011). In addition, systems that return an N-best list have employed metrics such as *Receiver Operator Characteristic* (*ROC*) and *Normalized Cross Entropy* (*NCE*) (Gandrabur et al., 2006), as well as *Weighted Semantic Error Rate* (*WSER*) and *Item-level Cross Entropy* (Thomson et al., 2008). The first three of these metrics evaluate confidence scores or overall correctness of the interpretations in an N-best list, while the last metric combines both aspects. Lee (2014) describes additional metrics that consider non-top dialogue-state hypotheses in an N-best list: Euclidean distance between a vector of the scores of all hypotheses and a 0 vector with 1 in the position of the correct hypothesis, average score of the correct hypothesis, and mean reciprocal rank of the correct hypothesis. However, with the exception of the last metric, none of these metrics takes into account the ranking of correct interpretations as done by the *NDCG@K* measure, and none of the metrics considers same-score interpretations as done by our *NDCG@K* and *FRecall@K* measures. In addition, our Interpretive experiment and its correlation-based evaluation differ from typical experiments in the literature.

7. Conclusion and future work

We have offered a numerical anytime mechanism for spoken language interpretation, focusing on descriptions, that considers multiple interpretations, and combines probabilities and scores from several sources, i.e., ASR, syntax, semantics and pragmatics. Within the latter, our mechanism integrates functions that represent lexical similarity as well as physical properties (viz colour, size, and topological and projective relations), which enables it to handle complex referring expressions.

We have conducted two types of evaluations: an Interpretive experiment, where we compared our system's understanding of written descriptions, focusing on ambiguous and imperfect descriptions, with that of trial participants; and a Generative experiment, where we tested our system's performance on open-ended spoken and transcribed descriptions. The results of our Interpretive experiment show that *Scusi?*'s understanding of (mostly) ambiguous and imperfect descriptions largely matches that of people, and when it does not, *Scusi?*'s interpretations are plausible. The results of our Generative experiment show that, as expected, *Scusi?*'s performance for written descriptions is considerably better than for spoken descriptions, and that *Scusi?* is able to overcome a significant proportion of ASR errors, and understand a substantial number of descriptions that contain unknown concepts.

We have considered two schemes for combining descriptive attributes, viz Multiplicative and Additive; and two sets of weights for prioritizing these attributes and the different interpretation stages, viz unity weights and weights learned by a steepest ascent hill climbing algorithm. Our results show that *WeightedScusi?*, which uses the weights learned by steepest ascent, outperforms *UnityScusi?*, which employs weights equal to 1, on input from the ASR. Our Generative trial indicates that ambiguous and imperfect descriptions are less frequent in naturally occurring descriptions than in our Interpretive experiment (Table 8). Hence, although the Additive scheme slightly outperforms the Multiplicative scheme in the Interpretive experiment, we consider the Multiplicative scheme to be superior, because it outperforms the Additive scheme in the larger, more realistic Generative experiment.

Our approach was motivated by the need to be able to function in the real word, which involves handling ambiguous and imperfect utterances, and dealing with unknown vocabulary and grammar. As shown in our Interpretive experiment in particular, *Scusi?* is able to deal with ambiguous and imperfect descriptions. In addition, our Generative trial demonstrates that *Scusi?* can handle a significant number of nouns and noun modifiers that are unknown to *Scusi?* due to its limited vocabulary or geometric knowledge, or as a result of ASR error. However, as seen in our Generative trial, additional phenomena must be addressed in order to enable *Scusi?* to act in the real world, such as coping with disjointed speech (Komatani et al., 2014) and speech disfluencies (Lison and Kruijff, 2009; Germesin et al., 2008; Gordon et al., 2011), and dealing with ungrammatical input (Lison and Kruijff, 2009), out-of-grammar input (Komatani et al., 2008) and unexpected utterances (Moratz and Tenbrink, 2006). To address the last three phenomena, we propose to extend our current approach by identifying regions of a description that exceed *Scusi?*'s capabilities, and assessing the effect of removing or replacing these regions both on the integrity of the remaining information, and on *Scusi?*'s ability to process this information. Indeed, we have conducted preliminary experiments in this vein, where we employed generic words to replace words whose PoS had a low probability according to *Scusi?*'s language model. For instance, "the battle *played* in the microwave" (where the PoS of "played" is VB) was replaced with "the battle *thing* in the microwave", which enabled *Scusi?* to find the intended referent (Kim et al., 2013).

Our mechanism suffers from three main limitations, which will be addressed in future work.

- As mentioned in Section 4, several operating parameters were manually set, in particular the number of iterations, the thresholds for not expanding a node, and the buffer size $k_{\max}$ (Section 2). These parameters interact with the weights associated with the different interpretation stages and descriptive attributes (Section 4). In the future, we propose to investigate the application of machine learning and search techniques to learn these parameters.
- Several approaches have been proposed to improve ASR performance, viz using contextual information (Lison and Kruijff, 2008), considering advisors (Gordon et al., 2011), or employing a robust grammar (Lison and Kruijff, 2009). In the future, we will investigate the impact of these approaches on *Scusi?*'s performance.
- The spatial reasoning capabilities of our system may be enhanced by (1) considering additional relations (e.g., ordinals) and groups of objects, and complex positional relations; (2) incorporating different frames of reference (Trafton et al., 2005; Moratz and Tenbrink, 2006; Liu et al., 2010) and additional factors that influence the use of positional prepositions (Coventry and Garrod, 2004); (3) employing more precise bounding boxes (Section 4); and (4) investigating colour–distance schemes that have greater psychological validity than CIE94. However, these enhancements must be subject to an analysis of their potential benefit, compared to the advantages of the above-mentioned approach which circumvents phenomena that are beyond *Scusi?*'s knowledge.

Finally, in order to incorporate *Scusi?* into a Spoken Dialogue System, we will re-calibrate *Scusi?*'s scores as described in Section 5.3.3, and augment *Scusi?*'s interpretations with information obtained from previous dialogue turns (Lee, 2014; Williams et al., 2013).

Acknowledgment

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

Appendix A. Estimating intrinsic scores

In this section, we describe functions that model how well the lexical item, colour and size of a candidate object match the corresponding requirements in its parent UCG, and map the result to the (0,1] space.

Lexical item. We employ the Leacock and Chodorow (1998) WordNet-based similarity measure, denoted *LC*, to compute the similarity between the head noun in a specification (u_{lex}) and a noun commonly used to specify object k (k_{lex}). The *LC* similarity score, denoted S_{LC}, which indicates how suitable u_{lex} is for designating object k, is mapped to the [0, 1] range as follows:

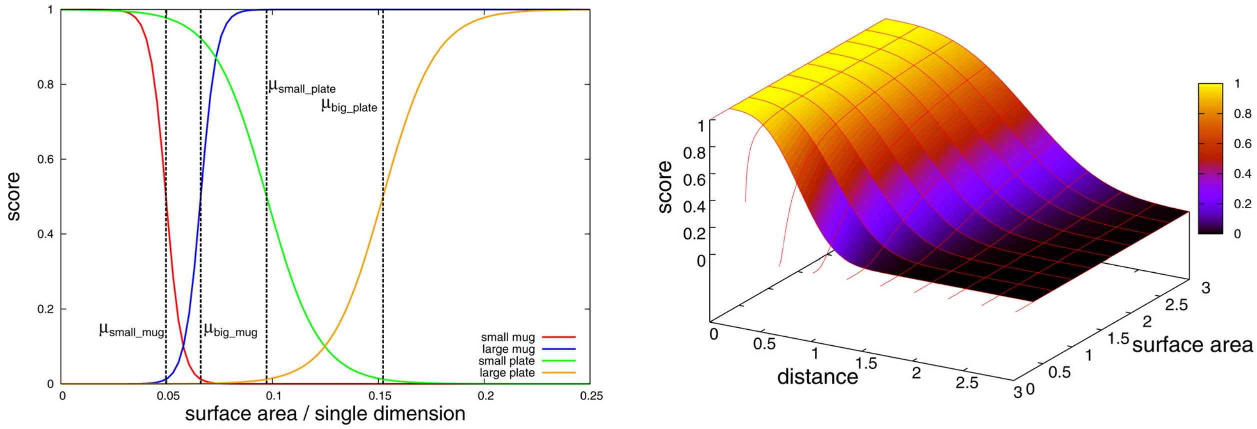
$$\text{Sc}(u_{\text{lex}}|k) = \max\left\{\epsilon, \frac{S_{LC}(u_{\text{lex}}, k_{\text{lex}}) - S_{\min}}{S_{\max} - S_{\min}}\right\}, \tag{13}$$

where $S_{\max}$ is the highest possible *LC* score, and $S_{\min}$ is the lowest *acceptable LC* score. This is the lowest score such that two objects are still considered similar, e.g., between “drawers” and “chest”. Any score below $S_{\min}$ is assigned a very small value ϵ. Such a value, rather than 0, still allows a candidate object to be considered – a useful behaviour when faced with speech recognition errors.

Although the *LC* measure yielded the best results among those in Pedersen et al. (2004), its similarity scores sometimes leave something to be desired, possibly due to the hierarchical nature of WordNet-based similarity scores. For instance, different objects have different similarity scores with the noun “thing”; “dinner table” and “desk” have the same similarity score with “table”; and sometimes completely different objects yield a higher similarity score than similar objects, e.g., S_{LC}(“shoe”,“table”) = 0.47, while S_{LC}(“chest”,“drawers”) = 0.35.

Colour. We have chosen the CIE94 (L, a, b) colour space, which has been experimentally shown to be approximately perceptually uniform (i.e., a small change in colour yields a small change in perception) (Rubner et al., 2001).[8] The L

[8] CIE94 solves some of the problems of CIE76, which was used in Makalic et al. (2008). For an introduction to this method, see http://en.wikipedia.org/wiki/Color_difference#CIE94.



(a) Score of size for small mugs and plates, and big mugs and plates

(b) Score of $near(x, y)$ as a function distance between x and y and surface area of the larger object

Fig. 7. Estimation of size and distance scores. (For interpretation of the references to colour in the text, the reader is referred to the web version of the article.)

coordinate represents brightness ($L = 0$ denotes black, and $L = 100$ white), a represents a range between green ($a < 0$) and red ($a > 0$), and b a range between blue ($b < 0$) and yellow ($b > 0$). We employ the following inverse linear conversion to map the CIE94 distance between a specified colour u_{col} and the colour of a candidate object k_{col} to a value in the [0, 1] range which reflects how well the colour stated in u_{col} matches the colour of object k:

$$\text{Sc}(u_{\text{col}}|k) = \max\left\{\epsilon, 1 - \frac{\max\{CIE94(u_{\text{col}}, k_{\text{col}}), CIE94(k_{\text{col}}, u_{\text{col}})\}}{CIE94_{\max}}\right\}, \tag{14}$$

where $CIE94(x, y)$ is the distance between the (L, a, b) coordinates of x and y, and $CIE94_{\max}$ is the maximum colour distance. Perceptual uniformity is a desirable property that enables CIE94 to yield more accurate colour distance measures overall than those obtained by simple metrics, such as the Euclidean distance in the RGB or HSV colour spaces. However, CIE94 suffers from two shortcomings: (1) it is asymmetric, i.e., in general $CIE94(x, y) \neq CIE94(y, x)$, because the coordinates of the "reference colour" are given more prominence; and (2) it does not handle well differences in saturation (e.g., pale blue is considered very different from blue). We overcome the first shortcoming by using the maximum of the distance between two colours, which yields the lowest match score (we also experimented with average and minimum, which yielded inferior results). As for lexical matches, to avoid completely invalidating a bad colour match, we employ ϵ as the lowest score.

Size. Intuitively, a request for a big object should match object instances whose size is greater than the mean, and a request for a small object should match object instances whose size is smaller than the mean. We model this behaviour by using two sigmoid functions to estimate how well the designation "big" or "small" suits an object of size k_{size} (the calculations for individual dimensions, viz length, width and height, are analogous):

$$\text{Sc}(u_{\text{size}} \in \text{big}|u_{\text{lex}}, k) = [1 + e^{\tau(k_{\text{size}} - \mu_{\text{big}})/\sigma}]^{-1} \tag{15}$$

$$\text{Sc}(u_{\text{size}} \in \text{small}|u_{\text{lex}}, k) = 1 - [1 + e^{\tau(k_{\text{size}} - \mu_{\text{small}})/\sigma}]^{-1}, \tag{16}$$

where μ and σ are the mean and standard deviation respectively of the surface area of objects of a particular type (e.g., balls or computer monitors), calculated from size measurements of household objects obtained from catalogues;[9] $\mu_{\text{big}} = \mu + \min\{\sigma, \mu/2\}$ and $\mu_{\text{small}} = \mu - \min\{\sigma, \mu/2\}$ are the means of the "big" and "small" distributions respectively of objects of a particular type ($\mu/2$ is used when the size distribution has a larger variance than μ); and τ (currently set to 0.2) defines the slope of the sigmoid function.

Fig. 7(a) shows the sigmoid functions for small and big mugs (red and blue lines) and small and big plates (green and orange lines), where the vertical bars depict μ_{big} and μ_{small} for mugs and plates. For instance, when a big plate is

[9] We use surface area, as it is less sensitive than volume to low values of individual dimensions.

specified, the “big” (orange) sigmoid assigns a score of 0.5 to candidate plates whose size is $\mu_{\text{big\_plate}}$. Plates with a size larger than $\mu_{\text{big\_plate}}$ have a higher score, which gradually approaches 1, while the score of plates that are smaller than $\mu_{\text{big\_plate}}$ gradually decreases to ϵ. Requests for small objects are treated analogously.

The suitability of a particular size term for describing the size of a candidate object depends on the synonymy between the term commonly used to refer to this object and the term given in the description, and on the size of other candidate objects of the same type. To illustrate the first aspect, consider the description “the small chair”. Since a stool may be synonymous with a chair, when matching a particular stool to the given description, *Scusi?* compares the size of this stool to that of a small chair, rather than to the size of a small stool. To illustrate the second aspect, consider the description “the bigger mug” uttered in a room with several small mugs and one average sized mug. At present, our approach does not differentiate between positive, comparative or superlative adjectives. Thus, for this description, bigger objects are assigned higher scores than smaller objects of the same type, regardless of their absolute size, and if none of the candidate objects are big, the resultant scores will be low (with the biggest mug having the highest score). This issue may be addressed by using adjective type to adjust the resultant scores. It is worth noting that our approach differs from that of Makalic et al. (2008) and Mitchell et al. (2011), who consider only relative size for understanding and generating size descriptors respectively. That is, they compare the size of a referent only to the sizes of distractors. Although this works well for comparative and superlative adjectives, it fails to determine whether a referent intrinsically matches a required size.

Appendix B. Estimating positional scores

In this section, we describe functions that model positional relations represented by the following topological and projective prepositions (Kelleher and Costello, 2008; Coventry and Garrod, 2004): *on, in, near, at* and *in/at/near the center/corner/edge/end of* (topological); and *in front/back of, behind*, *to the left/right of*, *above* and *under* (projective). Our approach for the representation of *near* and projective relations generally resembles Kelleher and Costello’s (2008), but it is applied to arbitrary household objects in realistic settings. Additional differences required to handle these situations are noted in the discussions for the individual relations.

At present, we assume that all the objects are rigid, and can be represented by a bounding box (Skubic et al., 2004) – the (x, y, z) coordinates of an object are the coordinates of the corner of its bounding box that is closest to the origin (0, 0, 0). We make the further simplifying assumption that the bounding boxes are aligned with the xy, yz and xz planes that match the orientation of the ceiling/floor and walls of a room containing the objects. These assumptions impose limitations for objects that are not perfect boxes (e.g., balls, plates, chairs), and objects that are askew in relation to the coordinates of the room. However, these limitations do not detract from the general validity of our approach, which may be refined with more precise bounding boxes. Like Kelleher and Costello (2008), we consider the presence of intervening objects between a candidate referent and a landmark when computing our *near* function, and the topological and projective functions that employ the *near* function. However, unlike Kelleher and Costello, we also take into account the size of the referent and the landmark.

B.1. Topological relations

The appropriateness of a topological specification is estimated on the basis of physical coordinates and dimensions of referents and landmarks.

On. The following description pertains to objects on the horizontal plane. In order to apply it to a vertical plane (e.g., “the picture *on* the wall”), the coordinates are rotated. A candidate referent k_r is *on* a landmark k_l if $k_{rz} = k_{lz} + k_{lh}$, where k_{rz} and k_{lz} are the z coordinates of the referent and the landmark respectively, and k_{lh} is the height of the landmark (that is, the referent must be *on top* of the landmark). In addition, we posit that the score assigned to a referent on a landmark is proportional to the degree of their overlap in the horizontal plane. If the above constraint is not satisfied or there is no overlap between the referent and the landmark, Sc(*Location _on*(k_r, k_l)) is set to a low score ϵ. These requirements yield the following formulation for the score Sc(*Location _on*(k_r, k_l)):

$$\mathrm{Sc}(Location\_on(k_r, k_l)) = \begin{cases} \dfrac{A_{xy}(k_r, k_l)}{\min\{A_{xy}(k_r), A_{xy}(k_l)\}} & \text{if } \{k_{rz} = k_{lz} + k_{lh} \;\&\; A_{xy}(k_r, k_l) > 0\} \\ \epsilon & \text{otherwise,} \end{cases} \tag{17}$$

where $A_{xy}(k)$ is the area of (the bounding box of) object k in the xy plane (horizontal surface), and $A_{xy}(k_r, k_l)$ is the overlapping horizontal area between k_r and k_l. For example, given the description "the book on the desk" in a context where `book1` is wholly on `desk3`, while `book2` is overhanging the side of `desk1`, *Location _on*(book1, desk3) will have a score of 1, while *Location _on*(book2, desk1) will have a lower score.

In/Inside. A candidate referent k_r is *in* a landmark k_l if its x and y coordinates are within the x and y coordinates of the landmark. In addition, the score assigned to a candidate referent k_r being inside a landmark k_l is proportional to the volume shared by the bounding boxes of k_r and k_l (one object could be partially inside another). Similarly to Eq. (17),

$$\mathrm{Sc}(\mathit{Location\_in}(k_r, k_l)) = \begin{cases} \dfrac{V(k_r, k_l)}{\min\{V(k_r), V(k_l)\}} & \text{if } \{k_{rx} > k_{lx} \text{ and } k_{rx} + k_{rw} < k_{lx} + k_{lw} \ \& \\ & \quad k_{ry} > k_{ly} \text{ and } k_{ry} + k_{rd} < k_{ly} + k_{ld}\} \\ \epsilon & \text{otherwise,} \end{cases} \tag{18}$$

where k_{*x}, k_{*y}, k_{*w} and k_{*d} are the x and y coordinates, and the width and depth of referent/landmark k respectively; $V(k)$ is the volume of (the bounding box of) object k; and $V(k_r, k_l)$ is the overlapping volume between k_r and k_l. For example, given the description "the mug inside the box", a mug that is wholly contained within a box yields a higher score than a mug whose top exceeds the top of a box.

Near/Far from. We use a base 2 negative exponential function to model the score assigned to a candidate referent being near a landmark (the score assigned to a referent being *far from* a landmark is the complement of the score of it being near the landmark):

$$\mathrm{Sc}(\mathit{Location\_near}(k_r, k_l)) = \max\left\{\epsilon, 2^{-\tau D(k_r,k_l)^{\alpha}/S_{\max}^{\beta}} \cdot (1 - \mathrm{Sc}(\mathit{occluded}(k_l)))\right\}, \tag{19}$$

where $D(k_r, k_l)$ is the distance between the referent and the landmark, S_{max} is the surface area of the larger object[10]; $\mathrm{Sc}(\mathit{occluded}(k_l))$ represents the extent to which an object between the referent and the landmark occludes the landmark from the speaker's view; and τ, α and β are parameters that define the shape of the negative exponential function. These parameters were set to $\tau = 1$, $\alpha = 3$ and $\beta = 1$, so that the function is more forgiving of large objects than small ones. This is necessary to handle examples such as "the key near the phone" as opposed to "the key near the desk", where the key may be farther from the desk than from the phone, and still be considered near the desk. Fig. 7(b) depicts the base 2 negative exponential function; note the gentler slope on the far side of the curve, which corresponds to objects with large surface areas.

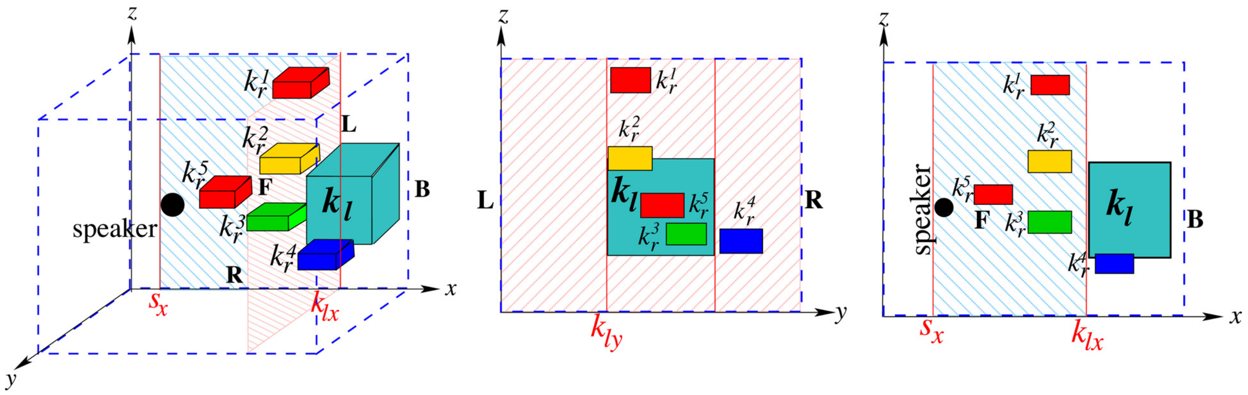
At. This is a vague positional relation which combines *on*, *in* and *near*, such that

$$\begin{aligned} &\mathrm{Sc}(\mathit{Location\_at}(k_r, k_l)) \\ &\quad = \max\{\epsilon, \mathrm{Sc}(\mathit{Location\_on}(k_r, k_l)), \mathrm{Sc}(\mathit{Location\_in}(k_r, k_l)), \mathrm{Sc}(\mathit{Location\_near}(k_r, k_l))\}. \end{aligned} \tag{20}$$

In/at/near the center/corner/edge/end of. In principle, like *on*, these positional relations may apply to landmarks in the horizontal or the vertical plane. However, at present our implementation covers only the horizontal plane.

The *center* specifier can only be *on* a landmark, while the other specifiers may be *on* or *off* the landmark. For instance, "the ball *near the corner of* the table" may be on or off the table, while "the ball *near the center of* the table" can only be on the table (currently we do not consider the center of a volume). We employ a combination of *on* and *near* to estimate the score of a candidate referent being in or near the *center* of a landmark. That is, we first check if the referent is *on* the landmark, and if so, estimate the score of the referent being *near* a 1 cm^2 square in the center of the landmark. The *corner/edge/end* specifiers are implemented using *near*, where the landmark is further narrowed down as follows: in the *off* version for a hollow object (e.g., a table) and the *on* version, the top of the bounding box of the landmark has four *corners*, each comprising a square of 1 cm^2 surface area. Similarly, an oblong object normally has four *edges* and two *ends*, each 1 cm wide ("the end of the table" normally refers to the narrow end, but a square table has four ends). In the *off* version of a convex object (e.g., a box), the bottom of the bounding box of the object has four additional *corners* and *edges*, and two additional *ends*.

[10] As for size, we use surface area. We also experimented with the surface area of the landmark, but the results were unintuitive when the landmark was smaller than the referent.



(a) 3D (b) Projection in the yz plane (c) Projection in the xz plane

Fig. 8. Illustration of *in front of* (off the landmark).

B.2. Projective relations

Like Kelleher and Costello (2008), we adopt a viewer-centered frame of reference for these relations.

In front/back of / Behind / To the left/right of. As above, with the exception of *behind*, these positional relations may be *on* or *off* the landmark.

The *on* variants of these specifiers are similar to *in/at/near the edge of*, but rather than selecting any edge, we select the appropriate one depending on the specifier and the speaker's position. Specifically, for *in front of*, the edge of the landmark that is closest to the speaker is chosen, for *in back of*, the edge that is farthest from the speaker is used, and so on. In addition, like for *center*, the object must be *on* the landmark.

We illustrate the implementation of the *off* variants by describing *in front of*, assuming that the speaker's x coordinate is lower than that of the landmark and the referent (Fig. 8(a)). The other specifiers are symmetrical, and are modeled by rotating the position of the speaker. Like Kelleher and Costello's (2008) model of these projective relations, ours is based on the model described by Logan and Sadler (1996). However, Kelleher and Costello do not take into account the dimensions of the referent and the landmark, and do not model the spatial overlap between them. For a referent to be considered in front of a landmark, it should (1) be between the speaker and the landmark in the xy plane (k_r^1, k_r^2, k_r^3 and k_r^5, Fig. 8(a)), and (2) overlap with or be very near the landmark in the yz plane (k_r^2, k_r^3, k_r^4 and k_r^5, Fig. 8(b)). In addition, the score assigned to objects that are near the landmark (k_r^2, k_r^3 and k_r^4, Fig. 8(c)) is higher than the score assigned to objects that are farther away (k_r^1 and k_r^5). These requirements are implemented as follows:

$$\begin{aligned}&\mathrm{Sc}(\mathit{Location\_inFrontOf\_Off}\ (k_r, k_l))\\&\quad = \mathrm{Sc}(\mathit{near}(k_r, k_l) | \mathit{inFrontRangeOf\_Off}(k_r, k_l)) \cdot \mathrm{Sc}(\mathit{inFrontRangeOf\_Off}(k_r, k_l)),\end{aligned} \tag{21}$$

where

$$\begin{aligned}&\mathrm{Sc}(\mathit{inFrontRangeOf\_Off}\ (k_r, k_l))\\&\quad = \begin{cases} \epsilon & \text{if } \{k_{rx} + k_{rw} \leq s_x \text{ or } k_{rx} \geq k_{lx}\} \\ 0.1 + 0.9 \cdot \dfrac{A_{yz}(k_r, k_l)}{\min\{A_{yz}(k_r), A_{yz}(k_l)\}} & \text{else if } A_{yz}(k_r, k_l) > 0 \\ 0.1 \cdot \max\{\mathrm{Sc}(\mathit{near}(k_r, k_l)), \mathrm{Sc}(\mathit{near}(k_l, k_r))\} & \text{otherwise} \end{cases}\end{aligned} \tag{22}$$

where s_x, k_{rx} and k_{lx} are the x coordinates of the speaker, the referent and the landmark respectively; k_{rw} is the width of the referent; and $A_{yz}(k_r, k_l)$ is the overlapping area between k_r and k_l in the yz plane. Thus, the first condition in Eq. (22) demands that the referent not be completely behind the speaker, and that the front of the referent be between the speaker and the landmark (part of the referent can overlap with the landmark, e.g., "the box in front of the table").

This condition invalidates k_r^4, which is not in front of k_l. The second condition above is addressed by the rest of Eq. (22): the "else if" part stipulates that the score of referents that overlap with the landmark on the *yz* plane is a function of the area of overlap, and is capped at a minimum threshold (set at 0.1). This yields a score of 1 for k_r^3 and k_r^5, and a lower score for k_r^2. The "otherwise" part allows the inclusion of referents that do not overlap with the landmark in the *yz* plane, but are very close to it, capping their maximum score at the threshold. This function resembles the projective function employed by Moratz and Tenbrink (2006) in the sense that the acceptance area is combined with a weighting scheme (which in our case is distance based). In our example, this function yields the highest score for k_r^3, followed by k_r^2 and k_r^5.

To implement the other specifiers, the speaker's perspective is simply rotated as indicated in Fig. 8(b) and (c): for *to the left of*, the speaker is moved to position **L**, for *to the right of* to position **R**, and for *behind* to position **B**. Note that these new speaker positions are at the walls, as any object between the landmark and the wall satisfies the given requirement, with referents near the landmark having a higher score.

Above/Under. In contrast to the other projective relations, *above* and *under* do not depend on the distance between the referent and the landmark, and are not influenced by the presence of intervening objects. For example, it is reasonable to refer to a ceiling lamp as "the lamp above the table". *Above* and *under* are implemented using a variant of Eq. (22) where the coordinates are rotated so that the speaker is respectively positioned directly under the ceiling or on the floor. However, for hollow objects, *under* demands that the referent be under the top of the landmark, i.e., the rotated k_{lx} in Eq. (22) is the bottom of the top of the landmark.

References

Black, A., Burger, S., Conkie, A., Hastie, H., Keizer, S., Lemon, O., Merigaud, N., Parent, G., Schubiner, G., Thomson, B., Williams, J., Yu, K., Young, S., Eskenazi, M., 2011. Spoken dialog challenge 2010: comparison of live and control test results. In: SIGDIAL2011 – Proceedings of the 12th SIGdial Meeting on Discourse and Dialogue, Portland, OR, pp. 2–7.

Coventry, K., Garrod, S., 2004. Saying, Seeing, and Acting: The Psychological Semantics of Spatial Prepositions. Psychology Press.

Dale, R., Reiter, E., 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. Cogn. Sci. 18 (2), 233–263.

DeVault, D., Sagae, K., Traum, D., 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In: SIGDIAL2009 – Proceedings of the 10th SIGdial Meeting on Discourse and Dialogue, London, United Kingdom, pp. 11–20.

Funakoshi, K., Nakano, M., Tokunaga, T., Iida, R., 2012. A unified probabilistic approach to referring expressions. In: SIGDIAL2012 – Proceedings of the 13th SIGdial Meeting on Discourse and Dialogue, Seoul, South Korea, pp. 237–246.

Gandrabur, S., Foster, G., Lapalme, G., 2006. Confidence estimation for NLP applications. ACM Trans. Speech Lang. Process. 3 (3), 1–29.

Gatt, A., van der Sluis, I., van Deemter, K., 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: ENLG07 – Proceedings of the 11th European Workshop on Natural Language Generation, Saarbrücken, Germany, pp. 49–56.

Germesin, S., Becker, T., Poller, P., 2008. Domain-specific classification methods for disfluency detection. In: Proceedings of Interspeech 2008, Brisbane, Australia, pp. 2518–2521.

Gordon, J.B., Passonneau, R.J., Epstein, S.L., 2011. Helping agents help their users despite imperfect speech recognition. In: Help Me Help You: Bridging the Gaps in Human–Agent Collaboration: Papers from the AAAI 2011 Spring Symposium, Palo Alto, CA, pp. 12–17.

Hüwel, S., Wrede, B., 2006. Spontaneous speech understanding for robust multi-modal human–robot communication. In: Proceedings of the COLING/ACL Main Conference Poster Sessions, Sydney, Australia, pp. 391–398.

Herrmann, T., Deutsch, W., 1976. Psychologie der Objektbenennung. Hans Huber.

Hirschman, L., 1998. The evolution of evaluation: lessons from the message understanding conferences. Comput. Speech Lang. 12, 281–305.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) 20 (4), 422–446.

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

Kelleher, J., Costello, F., 2008. Applying computational models of spatial prepositions to visually situated dialog. Comput. Linguist. 35 (2), 271–306.

Kim, S., Zukerman, I., Kleinbauer, T., Zavareh, F., 2013. A noisy channel approach to error correction in spoken referring expressions. In: IJCNLP2013 – Proceedings of the 6th International Joint Conference on Natural Language Processing, Nagoya, Japan, pp. 234–242.

Kleinbauer, T., Zukerman, I., Kim, S., 2013. Evaluation of the Scusi? spoken language interpretation system – a case study. In: IJCNLP2013 – Proceedings of the 6th International Joint Conference on Natural Language Processing, Nagoya, Japan, pp. 225–233.

Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., Lewin, I., 2001. Comparing grammar-based and robust approaches to speech understanding: a case study. In: EUROSPEECH 2001 – Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 1779–1782.

Komatani, K., Ikeda, S., Ogata, T., Okuno, H.G., 2008. Managing out-of-grammar utterances by topic estimation with domain extensibility in multi-domain spoken dialogue systems. Speech Commun. 50 (10), 863–870.

Komatani, K., Hotta, N., Sato, S., 2014. Restoring incorrectly segmented keywords and turn-taking caused by short pauses. In: IWSDS2014 – Proceedings of the International Workshop on Spoken Dialogue Systems, Napa, CA, pp. 27–38.

Krahmer, E., van Deemter, K., 2012. Computational generation of referring expressions: a survey. Comput. Linguist. 38 (1), 173–218.

Kruijff, G.-J., Lison, P., Benjamin, T., Jacobsson, H., Hawes, N., 2007. Incremental, multi-level processing for comprehending situated dialogue in human–robot interaction. In: LangRo'2007 – Proceedings from the Symposium on Language and Robots, Aveiro, Portugal, pp. 509–514.

Kruijff, G.-J.M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, I., Hawes, N., 2010. Situated dialogue processing for human–robot interaction. In: Christensen, H.I., Kruijff, G.-J.M., Wyatt, J.L. (Eds.), Cognitive Systems, Vol. 8 of Cognitive Systems Monographs. Springer, Berlin/Heidelberg, pp. 311–364.

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

Lee, S., 2014. Extrinsic evaluation of dialog state tracking and predictive metrics for dialog policy optimization. In: SIGDIAL2014 – Proceedings of the 15th SIGdial Meeting on Discourse and Dialogue, Philadelphia, PA, pp. 310–317.

Levelt, W., 1989. Speaking: from Intention to Articulation. MIT Press.

Lison, P., Kruijff, G.-J., 2008. Salience-driven contextual priming of speech recognition for human–robot interaction. In: ECAI 2008 – Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, pp. 636–640.

Lison, P., Kruijff, G.-J., 2009. Robust processing of situated spoken dialogue. In: Proceedings of the 32nd German Conference on Artificial Intelligence, Paderborn, Germany, pp. 241–248.

Liu, C., Walker, J., Chai, J.Y., 2010. Ambiguities in spatial language understanding in situated human robot dialogue. In: Dialog with Robots: Papers from the AAAI Fall Symposium, Arlington, VA, pp. 50–55.

Logan, G., Sadler, D., 1996. A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M., Garrett, M., Nadel, L. (Eds.), Language and Space. MIT Press, Cambridge, MA, pp. 493–529.

Makalic, E., Zukerman, I., Niemann, M., Schmidt, D., 2008. A probabilistic model for understanding composite spoken descriptions. In: PRICAI 2008 – Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, pp. 750–759.

Meena, R., Skantze, G., Gustafson, J., 2012. A data-driven approach to understanding spoken route directions in human–robot dialogue. In: Proceedings of Interspeech 2012, Portland, OR, pp. 226–230.

Mitchell, M., van Deemter, K., Reiter, E., 2011. Two approaches for generating size modifiers. In: ENLG2011 – Proceedings of the 13th European Workshop on Natural Language Generation, Nancy, France, pp. 63–70.

Moratz, R., Tenbrink, T., 2006. Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations. Spat. Cogn. Comput. Interdiscip. J. 6 (1), 63–107.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers, San Mateo, CA.

Pechmann, T., 1989. Incremental speech production and referential overspecification. Linguistics 27, 89–110.

Pedersen, T., Patwardhan, S., Michelizzi, J., 2004. WordNet::Similarity – measuring the relatedness of concepts. In: AAAI04 – Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, CA, pp. 25–29.

Pfleger, N., Alexandersson, J., Becker, T., 2003. A robust and generic discourse model for multimodal dialogue. In: Proceedings of the 3rd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Acapulco, Mexico.

Punyakanok, V., Roth, D., Yih, W., 2008. The importance of syntactic parsing and inference in semantic role labeling. Comput. Linguist. 34 (2), 257–287.

Ross, R., 2010. Putting things in context: situated language understanding for human–robot dialogue. In: Dialog with Robots: Papers from the AAAI Fall Symposium, Arlington, VA, pp. 103–108.

Rubner, Y., Puzicha, J., Tomasi, C., Buhmann, J.M., 2001. Empirical evaluation of dissimilarity measures for color and texture. Comput. Vis. Image Underst. 84 (1), 25–43.

Salmon-Alt, S., Romary, L., 2000. Reference resolution within the framework of cognitive grammar. In: INLG'2000 Workshop on Coherence in Generated Multimedia, Mitzpe Ramon, Israel.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., Brock, D., 2004. Spatial language for human–robot dialogs. IEEE Trans. Syst. Man Cybern. C 34 (2), 154–167 (Special Issue on Human–Robot Interaction).

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

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

Trafton, J.G., Cassimatis, N.L., Bugajska, M.D., Brock, D.P., Mintz, F.E., Schultz, A.C., 2005. Enabling effective human–robot interaction using perspective-taking in robots. IEEE Trans. Syst. Man Cybern. A: Syst. Hum. 35 (4), 460–470.

van Deemter, K., 2006. Generating referring expressions that involve gradable properties. Comput. Linguist. 32 (2), 195–222.

Wallace, C.S., 2005. Statistical and Inductive Inference by Minimum Message Length. Springer, Berlin, Germany.

Williams, J., Raux, A., Ramachandran, D., Black, A., 2013. The dialog state tracking challenge. In: SIGDIAL2013 – Proceedings of the 14th SIGdial Meeting on Discourse and Dialogue, Metz, France, pp. 404–413.

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