# Investigating the Dimensions of Spatial Language

**Adam Richard-Bollans**
University of Leeds, UK
mm15alrb@leeds.ac.uk

**Lucía Gómez Álvarez**
University of Leeds, UK
sc14lga@leeds.ac.uk

**Brandon Bennett**
University of Leeds, UK
b.bennett@leeds.ac.uk

**Anthony G. Cohn**
University of Leeds, UK
a.g.cohn@leeds.ac.uk

## Abstract

Spatial prepositions in the English language can be used to denote a vast array of configurations which greatly diverge from any typical meaning and there is much discussion regarding how their semantics are shaped and understood. Though there is general agreement that non-geometric aspects play a significant role in spatial preposition usage, there is a lack of available data providing insight into how these extra semantic aspects should be modelled. This paper is aimed at facilitating the acquisition of data that supports theoretical analysis and helps understand the extent to which different kinds of features play a role in the semantics of spatial prepositions. We first consider key features of spatial prepositions given in the literature. We then introduce a framework intended to facilitate the collection of rich data; including geometric, functional and conventional features. Finally, we describe a preliminary study, concluding with some insights into the difficulties of modelling spatial prepositions and gathering meaningful data about them.

**2012 ACM Subject Classification** Computing methodologies → Natural language processing; Computing methodologies → Lexical semantics

**Keywords and phrases** Lexical Semantics, Spatial Prepositions, Situated Dialogue

**Supplement Material** github.com/alrichardbollans/spatial-preposition-annotation-tool-blender

## 1 Introduction

Spatial prepositions in the English language can be used to denote a vast array of configurations which greatly diverge from any typical meaning and there is much discussion regarding how their semantics are shaped and understood. Though there is general agreement that non-geometric aspects play a significant role in spatial preposition usage, there is a lack of available data providing insight into how these extra semantic aspects should be modelled.

This paper is aimed at facilitating the acquisition of data that supports theoretical analysis and helps understand the extent to which different kinds of features play a role in the semantics of spatial prepositions. We first consider key features of spatial prepositions given in the literature. We then introduce a framework[1] intended to facilitate the collection of rich data; including geometric, functional and conventional features. Finally, we describe a preliminary study, concluding with some insights into the difficulties of modelling spatial prepositions and gathering meaningful data about them.

[1] github.com/alrichardbollans/spatial-preposition-annotation-tool-blender/




In Proceedings Speaking of Location 2019: Communicating about Space, Regensburg, Germany, September-2019.
Editors: K. Stock, C.B. Jones and T. Tenbrink (eds.);
Published at http://ceur-ws.org

## 2 Background

Regarding the names of the objects being discussed we use *figure* (also known as: target, trajector, referent) to denote the entity whose location is important e.g. 'the **bike** next to the house' and *ground* (also known as: reference, landmark, relatum) to denote the entity used as a reference point in order to locate the figure e.g. 'the bike next to the **house**'.

### 2.1 Semantic Complexity

Initial attempts to understand and model spatial language naturally focused heavily on geometry. However, as has been recognised in the past couple of decades, spatial constraints are not enough to fully characterise spatial prepositions [1, 5, 6, 10, 14]. The use of prepositions is determined by geometric, functional *and* conventional considerations, as evidenced in [5, 8, 10].

Talmy [25] introduced and highlighted the importance of 'force-dynamics' in language and cognition, considering the force interactions of objects as a primitive notion that pervades language through metaphor. This work inspired future researchers to pay more attention to the force interactions present; most notably in the investigations of [6, 10] which considered the interactions of geometry and functionality in spatial semantics, in particular highlighting that the functional control of the ground over the figure strongly influences preposition usage.

Garrod et al. [10] give the well-cited example that a pear may be considered as *in* a bowl when it is not even partially contained by the convex hull of the bowl — if it is sat on top of a pile of other pears in the bowl. We also see examples of this in our collected data. It has also been shown that the way objects are labelled and conceptualised affects preposition use. Coventry et al. [5] found that when given exactly the same scene of an object on a plate/dish, humans will describe the configuration as *in* when the 'plate' is labelled as a dish, and *on* when labelled as a plate. It is suggested that this is due to the affordances associated with the concepts 'plate' and 'dish'.

Following [22], we therefore believe that a full semantic account of spatial prepositions, particularly for *on* and *in*, ought to include distinct functional & conventional components in order to (1) be closely aligned with human usage and understanding and (2) aid automated interpretation and generation of spatial expressions.

### 2.2 Related Work

The work of Platonov & Schubert [21] is closely aligned with the current work. In order to create and test a computational model of spatial prepositions they also generate 3D scenes to annotate. The annotations are created via two separate tasks, a 'Truth-Judgement Task' and a 'Description Task'. In both tasks participants are shown screenshots of the scene. In the truth judgement task participants are asked if a preposition is fitting for two objects, e.g. 'Is the cube on the table?', and respond by selecting an option from 'Yes', 'Rather Yes', 'Uncertain', 'Rather No', 'No'. In the Description Task participants are given an object, by referring to the object label e.g. 'Where is Pencil 1?', and asked to provide a description of the location of the object. On the whole, despite some minor deficiencies[2], this represents a valuable groundwork for our framework, which we enrich by supporting tasks in 3D environments (instead of screenshots) where participants can navigate the scenes and select objects. We believe that such additional features may be crucial for the integration of certain

[2] Many annotations were obtained from blocks world environments or simple uncluttered indoor scenes.

pragmatic factors in different tasks as well as provide more flexibility for the exploration of borderline geometrical configurations. Moreover, Platonov & Schubert's framework does not seem to be intended for reuse by different projects.

There exist some wide-ranging annotated datasets, see [9, 27], however the annotations are not restricted to spatial language and many of the datasets are image-based so extracting meaningful features is extremely difficult.

Many experimental studies have been conducted over the past couple of decades into particular aspects of spatial prepositions; however these are either image-based [2, 7, 15, 17, 19] or in real environments [10, 20, 26] where feature extraction is difficult; or in very constrained/simple environments [8, 11, 12, 16].

Overall, we find that there is a lack of detailed geometric, functional and contextual data which hinders the capacity to properly investigate the semantic complexity of spatial prepositions and provide pragmatic analysis on how they are used to achieve communicative success.

## 3 Framework

Our framework[1] is built on the 3D modelling software Blender[3]. This software allows 3D scenes to be created, which can then be converted into annotation environments using the built-in game engine. Scenes are easy to create, simply taking the given scene template and populating it with objects, see the GitHub repository[1] for instructions. See Figure 1 for a screenshot of a scene from our study (some object labels have been added but were not visible for participants during the study). To allow for easy selection, objects in the scene are indivisible entities e.g. a table in the scene can be selected but not a particular table leg.


**Figure 1** Example Scene

Once a scene has been created, a python script is run in order to create distinct tasks from the scene. Two tasks that we used for our preliminary study are described below.

[3] `https://www.blender.org/`

## 3.1 Tasks

The framework initially provides two distinct tasks with which our preliminary study was conducted — a Selection Task and a Description Task. The Selection Task was designed to efficiently collect large amounts of data regarding the semantics, with minimal pragmatic considerations. The Description Task provides more focused data which is intended to aid pragmatic analysis and test models of figure selection. In both tasks, participants are given a first person view of a scene which they can navigate using the mouse and keyboard.

In the Selection Task participants are given a preposition on screen and asked to select all figure-ground pairs in the scene which fit the preposition. Once they have selected all pairs they believe to be admissible they are shown another preposition and asked to repeat the process. As the output of each selection is a figure, preposition and ground no post-processing is required to identify these. In our preliminary study, we limited this task to prepositions we believe to have a functional component along with those prepositions that seem to act as a geometric counterpart: 'in', 'inside', 'on', 'on top of', 'against', 'over', 'under', 'above' and 'below'.

In the Description Task objects are highlighted and participants are able to type in a spatial description of the object. In our preliminary study we asked participants to give descriptions of the object locations using a definite description, in the format *figure* + *preposition* + *ground* e.g. 'the guitar by the bookshelf'. We also allowed the use of multiple prepositions if the participant deemed it necessary e.g. 'the cup on the table near the lamp'. In order to increase the number of annotations containing prepositions with a functional component, while still allowing participants a choice of natural descriptions, we asked participants to only use the prepositions in the Selection Task plus 'to the right of', 'to the left of', 'in front of', 'behind', 'near', 'next to', 'at'.

## 3.2 Feature Extraction

The use of virtual 3D environments allows for the extraction of a wide range of features that would not be immediately available in real-world or image based studies.

### 3.2.1 Geometric Features

Geometric features (distance between objects, object size etc..) are relatively simple to extract. We made use of and adapted some existing code[4] for this purpose, see [21]. The geometric features extracted are all quantitative rather than qualitative and some simplifications have been made. For example, we measured *containment* as the proportion of the bounding box of the figure that is shared with the bounding box of the ground. This measure may be improved by refining how the shared volume is calculated (e.g. by using convex hulls rather than bounding boxes) or by distinguishing separate parts of the ground where overlaps with the figure are more important (e.g. the *containable inside* [4] of the ground). However, calculations involving convex hulls can become computationally expensive and automated demarcation of such salient parts of objects is a non-trivial task.

[4] `https://github.com/gplatono/SRP/tree/master/blender_project` Date Accessed: 22/11/2018

### 3.2.2 Functional Features

There are two particular functional notions that appear over and over in the literature: 'support' and 'location control'. It is often considered that 'on' expresses the functional notion of 'support' while 'in' expresses 'location control' [3]. We take 'support' to express that the ground impedes motion due to gravity of the figure, while 'location control' expresses that moving the ground moves the figure. By including these features in our data we hope to provide significant support for the theories of [6, 10]. Rather than attempting to formally define these notions, as in [13], following Sloman [23] we quantified these notions via *simulation* using Blender's built-in physics engine. To assess the degree to which a ground gives support to a given figure, we measure the change in vertical position of the highest point of the figure when the ground is removed from the scene. We then normalised this measure by dividing by the height of the ground.

For simplicity, we have used this support measure as a proxy for location control as it is easier to quantify and calculate from a scene. Further, the notions are conceptually quite similar and there are findings to suggest that the notions of containment and support are closely related in humans' conceptual framework [18].

### 3.2.3 Conventional Features

In order to assess the degree to which non-contextual features affect preposition use, we used ConceptNet [24] to gather object properties. ConceptNet is a large-scale relational knowledge base for commonsense knowledge, taking information from multiple crowd-sourced projects. Currently, we only considered a single commonsense feature — to discern whether the ground object is considered a *container*. To do this, for a given ground, we extracted the weight of the 'IsA' edge between the ground and the concept 'container'.

## 4 Preliminary Study

In this section we describe a preliminary study carried out with our framework and discuss the results. See the GitHub repository[1] for the collected data.

We created three separate virtual indoor environments which were each limited to around 60 household objects. Each scene contained a robot and for the Description Task the participants were asked to describe the objects in such a way that the robot would pick up the correct object in order to bring it to them. We chose to only highlight objects which were not uniquely identifiable in the scene by their name i.e. where there were multiple objects of the same type in the scene.

### 4.1 Session & Resulting Data

To carry out the study we hosted a session in one of our computing labs and invited participants from across the university campus to take part[5]. We asked participants to complete either the Selection Task or Description Task in one of our scenes. Participants were given a brief introduction to the study and instructions on how to complete the given task along with explanations of key terminology[6]. All participants were asked to create

[5] Ethics Approval Code: 271016/IM/216. Participants were given the incentive of free pizza.
[6] See GitHub repository[1] for instructions that were given.

**Table 1** Selection Task: Total number of selections made for each preposition & the number of distinct configurations selected

| Preposition | Above | Over | Below | Under | Against | In | Inside | On | On top of |
|---|---|---|---|---|---|---|---|---|---|
| Selections | 103 | 15 | 32 | 71 | 53 | 27 | 29 | 208 | 171 |
| Distinct Configurations | 83 | 14 | 30 | 65 | 46 | 15 | 15 | 149 | 91 |

annotations for ten minutes. 23 native English speakers participated, of whom 13 performed the Selection Task and 10 performed the Description Task.

For the Selection Task only minimal cleaning of the data was required. We removed one user that appeared to give selections at random (52 Selections). We also removed annotations repeated by the same user (153 Selections), annotations in which the same object was selected for figure and ground (19 Selections) and also annotations where a user had selected a pair of objects twice but reversing the figure-ground selection (11 Selections). Also, during the study participants stated that they had wanted to select the room to say there was objects 'in' or 'inside' it. We therefore removed annotations that we believe were intended for this purpose (4 Selections). Post-cleaning, we obtained 709 annotations in the Selection Task; see Table 1 for a breakdown of selections by preposition.

We obtained 245 annotations in the Description Task. We did not clean this data in any way but singled out 'simple' descriptions (consisting of a single figure, ground and preposition) to make analysis easier. We applied the off-the-shelf NLTK[7] Part-Of-Speech (POS) tagger, which labels each individual word with a POS tag, to find the simple descriptions. We added some hard-coded rules to deal with prepositional phrases of more than one word which are present in our data. This left some issues regarding more complex noun phrases e.g. 'the book on the square table' where the parser identifies three nouns — 'book', 'square' and 'table'. However, on the whole most of the simple phrases were identified and the phrases identified as 'simple' were indeed simple. There are 114 such 'simple' descriptions.

In the following sections we look at potential insights gained from the preliminary study, focusing on 'in' and 'on'.

## 4.2 Selection Task

### Containment & Containers

Apart from the exception of the bookcase, regardless of the geometric and functional relations, users annotating our scenes would only select 'in' for a pair of objects if the ground was a *container*[8]. We believe this sort of insight is potentially useful for systems aiming to generate realistic natural language expressions and in making pragmatic inferences related to possible descriptions of an object.

### Similarity of 'in' & 'on'

We found that there was significant overlap between 'in' and 'on'. This highlights some of the complexity of modelling this language; borderline scenarios are common and often more than one preposition can be used in a given situation. Out of the 15 distinct configurations

[7] `http://www.nltk.org/` Date Accessed: 05/11/18
[8] We call an object a container if there exists an 'IsA' edge between it and 'container' in ConceptNet.

**Table 2** Success of simple models

| | Proximity | | | | | Containment | |
|---|---|---|---|---|---|---|---|
| Preposition | On | On top of | In | Inside | Against | In | Inside |
| Correct | 25 | 8 | 6 | 1 | 2 | 6 | 1 |
| Incorrect | 27 | 4 | 1 | 1 | 1 | 1 | 1 |
| Unmatched | 3 | 2 | 4 | 0 | 0 | 4 | 0 |

that were selected as 'in', 12 of them were also selected as 'on'. We believe there were two main reasons for this: ambiguity of the ground's role as a container & the ground being a container but the relationship between figure and ground being typical for 'on'. For five of these configurations it is ambiguous whether the ground is a 'container' e.g. books were labelled as both 'in' and 'on' the bookcase/shelf. We believe this may partly be explained by synecdoche i.e. the bookcase is being used to refer to one of its particular shelves which the book is on. We also believe this is partly due to the fact that 'in' is generally used where the ground is a *container*, and it is somewhat ambiguous as to whether a bookcase is a container or not and therefore whether 'in' or 'on' is suitable. For another five of these configurations the ground is a container, but the relationship between the figure and ground is near the typical notion for 'on', for example 'banana' and 'bowl' in Figure 1.

### Location Control & Support

A key aspect that we wanted to analyse was the role of functional features. Support appears important to 'on' — the average value of support for 'on' was 0.72 compared to an average for all prepositions of 0.41. However, this measure by itself was not indicative of 'in', having an average value of 0.4. We believe, however, that this measure will be more informative in a more nuanced model that accounts for the distinct but related meanings, or *polysemes*, of 'in'. These results suggest that refinements of our data collection are necessary in order to make any significant claims.

The Selection Task relied on the thoroughness of participants in selecting all admissible configurations for each preposition, however we found that our scenes likely contained too many objects for this to be a reliable outcome for every participant. This hampered our efforts to provide any significant analysis.

## 4.3 Description Task

### Detailed Semantics & Communicative Success

Many computational models of spatial prepositions achieve some success when only accounting for relatively simple geometric features. It is evident that in order to locate items when given a spatial description (given that all the objects in the scene are known), one can usually achieve a high degree of accuracy without a detailed semantic model. We expect that, when given a locative expression, these sorts of models would often be sufficient to locate the figure.

Moreover, though prepositions may encode a large amount of information and the speaker's choice of preposition is dependent on many features, some often imply proximity and this alone can be useful in locating the figure object. We imagine that 'in','on','by','at' and 'against' expressions can often be successfully decoded by only considering proximity to the ground.

To test this we constructed simple models that would take a definite description as input and output an object from the scene. Given an expression '*figure* + *preposition* + *ground*' e.g. 'the book on the table', the models would first find all possible figure-ground pairs e.g. all possible books and tables in the scene. The Proximity Model would then output the figure object where the pair has the shortest distance. The Containment Model would output the figure object where the pair has the highest degree of containment. See Table 2 for the results of these models[9].

Note that due to the ambiguous nature of many of the given descriptions (discussed below), it is not possible to give an accurate measure of the success of these models. These results, however, indicate that simple strategies can often be successful.

Though in general simple solutions can be useful for the problem of figure selection in indoor environments, it is clear there are situations where more nuance in the model would be necessary. It is also not clear that such simple models would be robust enough to be reusable across different domains and tasks. Moreover, we expect that a more detailed picture is necessary for the purposes of language generation. However, we need to conduct further studies to clarify this.

### Pragmatics & Overcoming Ambiguity

We hoped that participants would give descriptions that were ambiguous (e.g. 'the book on the table' where there are two instances of a book on a table) that would be unambiguous in practice due to pragmatic considerations. Such pragmatic considerations may encapsulate a wide variety of world-knowledge, from relatively shallow features such as object occlusion to more task-dependent and commonsense factors such as proximity of an object to the speaker, as well as semantic knowledge, such as the acceptability of potential descriptions.

When analysing the 'simple' descriptions given by participants (containing only one preposition), even as a human, many appear genuinely ambiguous e.g. 'the notebook on the table' in Scene 2 (see Figure 1). This could be as a result of participants not being fully aware of all the objects in the scene, not having properly read the instructions or that the aim of the task was not made clear enough. Further, to discern pragmatic strategies used by participants it is necessary to focus on a specific pragmatic aspect as otherwise the problem is too varied and complex. As a result, we were unable to provide any meaningful pragmatic analysis. Further study is necessary and modifications need to be made in order to ensure the collected data is useful in future.

## 5 Conclusions

In this paper we have introduced our framework for collecting data on spatial prepositions and discussed a preliminary study conducted. By creating a framework that allows for the collection of rich data on spatial prepositions which includes geometric, functional and conventional features, we intend to begin a process of providing large-scale support for theoretical claims in the field, which will hopefully help not only to clarify their semantics but also to develop appropriate strategies to model them. Along these lines, our preliminary study has given some tentative insights and highlighted aspects of our data collection methodology that need further refining.

[9] Objects were unmatched where a suitable ground was not recognised.

## 6 Future Work

In order to build upon the current work, based on some of the discussion provided here, we are extending and improving upon the existing framework and running further studies. We have split the Selection Task into two separate tasks so that we can provide a more meaningful and robust analysis. One task shows participants a specific configuration and asks which prepositions could be used to describe it. This means that we are less reliant on participants being thorough and also can collect non-instances of a preposition. A second task shows participants a ground object as well as potential figure objects and asks them to select the figure object that best fits a given preposition. This will allow us to directly assess how features influence the typicality of spatial prepositions. The data collected in these first two tasks will be used to inform a prototype dialogue system that will be implemented in a description task. This task will take the form of a game, such as the user providing descriptions to collect objects in a scene for some task where they are given a score based on their performance. Our prototype system will be used to interpret the user's description and if it guesses correctly, the object will disappear from the scene and be added to the users inventory. Finally, in order to facilitate larger ongoing studies, we have created an updated web version of our framework which is currently online[10].

## References


1 John A Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071, 2010.

2 Melissa Bowerman and Erik Pederson. Topological relations picture series. In *Space stimuli kit 1.2*, page 51. Max Planck Institute for Psycholinguistics, 1992.

3 Laura Carlson, Emile Van der Zee, and Emile Zee. *Functional features in language and space: insights from perception, categorization, and development*, volume 2. Oxford University Press on Demand, 2005.

4 Anthony G Cohn, David A Randell, and Zhan Cui. Taxonomies of logically defined qualitative spatial relations. *Journal of human-computer studies*, 43(5-6):831–846, 1995.

5 Kenny R. Coventry, Richard Carmichael, and Simon C. Garrod. Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, 11(4):289–309, 1994.

6 Kenny R Coventry and Simon C Garrod. *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press, 2004.

7 Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. The Interplay between Geometry and Function in the Comprehension of Over, Under, Above, and Below. *Journal of Memory and Language*, 44(3):376–398, 2001.

8 Michele I Feist and Derdre Gentner. On plates, bowls, and dishes: Factors in the use of English IN and ON. In *Proc 20th annual meeting of the cognitive science society*, pages 345–349, 1998.

9 Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao, Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. A Survey of Current Datasets for Vision and Language Research. *arXiv:1506.06833 [cs]*, 2015.

10 Simon Garrod, Gillian Ferrier, and Siobhan Campbell. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189, September 1999.

11 Dave Golland. *Semantics and Pragmatics of Spatial Reference*. PhD Thesis, University of California, Berkeley, USA, 2013.


[10] `http://adamrichard-bollans.co.uk/spatial_language_project.html`

**12** Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *Proc IROS*, pages 1640–1647. IEEE, 2013.

**13** Maria M. Hedblom, Oliver Kutz, Till Mossakowski, and Fabian Neuhaus. Between Contact and Support: Introducing a Logic for Image Schemas and Directed Movement. In *Proc IAAI*, volume 10640, pages 256–268. Springer, 2017.

**14** Annette Herskovits. *Language and spatial cognition*. Cambridge University Press, 1987.

**15** John Kelleher, Colm Sloan, and Brian Mac Namee. An investigation into the semantics of English topological prepositions. *Cognitive Processing*, 10(S2):233–236, September 2009.

**16** John D. Kelleher and Fintan J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.

**17** John D. Kelleher, Geert-Jan M. Kruijff, and Fintan J. Costello. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In *Proc ACL*, pages 745–752. Association for Computational Linguistics, 2006.

**18** Jean M Mandler. How to build a baby: II. Conceptual primitives. *Psychological review*, 99(4):587, 1992.

**19** David M. Mark and Max J. Egenhofer. Topology of prototypical spatial relations between lines and regions in English and Spanish. In *Proc Auto Carto*, pages 245–254, 1995.

**20** Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006.

**21** Georgiy Platonov and Lenhart Schubert. Computational Models for Spatial Prepositions. In *Proc 1st International Workshop on Spatial Language Understanding*, pages 21–30, 2018.

**22** Adam Richard-Bollans. Towards a Cognitive Model of the Semantics of Spatial Prepositions. In *ESSLLI Student Session Proceedings*. Springer, 2018.

**23** Aaron Sloman. Aiming for More Realistic Vision Systems. Technical Report COSY-TR-0603, University of Birmingham, 2006.

**24** Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, pages 3679–3686, 2012.

**25** Leonard Talmy. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100, 1988.

**26** Thora Tenbrink, Elena Andonova, Gesa Schole, and Kenny R. Coventry. Communicative Success in Spatial Dialogue: The Impact of Functional Features and Dialogue Strategies. *Language and Speech*, 60(2):318–329, 2017.

**27** Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual Question Answering: A Survey of Methods and Datasets. *arXiv:1607.05910 [cs]*, 2016.