Grounding, Justification, Adaptation: Towards Machines That Mean What They Say

David Schlangen
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
david.schlangen@uni-bielefeld.de

Abstract

Meaningful language use rests on the *grounding* of the language in the non-linguistic world and in the practices of language users. This grounding is built up and maintained in interaction, through *Conversational Grounding*, which is the interactive process with which interlocutors build mutual understanding; *Justification*, the ability to explain and provide reasons for one's language use; and *Adaptation*, the ability to accept corrections and adapt future language use accordingly. We outline a model of grounded semantics that combines perceptual knowledge (how to visually identify potential referents of terms; realised as classifiers taking visual information as input) and taxonomic knowledge (covering lexical relations such as hyponymy and hypernymy), and we sketch a proof-of-concept implementation of a dialogue system that realises the interactional skills that ground this knowledge.

1 Introduction

Computer systems that process natural language input and produce natural language output are becoming ever more common and ever more capable. So-called "intelligent personal assistants" built into mobile phones are already serving real customer needs (e.g., providing verbal access to the user's calendar), and current research systems show impressive results on tasks like image captioning (given an image, produce a textual description of its content). And yet, there is a strong sense in which these system do not *mean* anything with their use of natural language. Why is that so?

We propose that meaningful language use rests on the *grounding* of the language: in the non-linguistic world; in itself, among the parts of the language; and in the practices of the community of language users. These are, at a least to a certain degree, complementary aspects, as Hilary Putnam (1973) pointed out with the claim that someone who (like him) cannot reliably tell an elm from a beech tree would still mean the same with *elm* as someone who can. Marconi (1997) uses this observation to motivate a model of what he calls *lexical competence* that separates *referential competence*—the competence to identify actual referents, which Putnam claims to lack with respect to elms—and *inferential competence*, which uses semantic knowledge to place meanings in relation to other meanings (here, for example, the relation of hyponymy between *elm* and *tree*).

This grounding is not static, however, but rather is built up and maintained in interaction, through *Conversational Grounding*, which is the interactive process with which interlocutors build mutual understanding; *Justification*, the ability to explain and provide reasons for one's language use; *Adaptation*, the ability to accept corrections and adapt future language use accordingly.

Our aim in this paper is to outline a model of semantic competence that can address these desiderata: That it explains what kind of discriminatory power constrains meaningful language use, and that this power is acquired, defended and adapted in interaction. Its basis is a "two dimensional" model of lexical knowledge. In this model, one dimension captures '*know-how*' such as the knowledge required to pick out the intended referent in a visually presented scene, and the other captures more semantic knowledge ('*know-that*') that enables inferences, but can also, as we show, support visual reference resolution. (This distinction is inspired by that between *referential* and *inferential* lexical competence made by Marconi (1997), but further generalised. The visual-grounding model builds on (Kennington and Schlangen, 2015). See discussion below.) Both kinds of knowledge can be trained from interaction data. The lexical representations are used to compose meanings of larger phrases. This composition process is transparent (compared to composition in distributional semantics, as discussed below), and hence is accessible for inspection and correction.

To make our proposal concrete, and to investigate the utility of interaction that has the system's own semantic competence as its topic, we implemented an interactive system that tries to resolve references to objects in images and can accept corrections and provide explanations. Figure 1 shows an excerpt of an actual interaction with the system of a naive first-time user. (Image sources are credited in Section 7.1 below.)

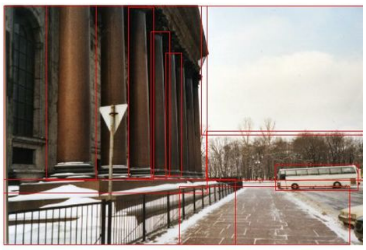
2 Related Work

The idea of connecting words to what they denote in the real world via perceptual features goes back at least to Harnad (1990), who coined "The Symbol Grounding Problem": "*[H]ow can the semantic interpretation of a formal symbol system be made* intrinsic *to the system, rather than just parasitic on the meanings in our heads?*" The proposed solution was to link 'categorial representations' with "learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections".

This suggestion has variously been taken up in computational work. An early example is Deb Roy's work from the early 2000s (Roy et al., 2002; Roy, 2002; Roy, 2005). In (Roy et al., 2002), computer vision techniques are used to detect object boundaries in a video feed, and to compute colour features (mean colour pixel value), positional features, and features encoding the relative spatial configuration of objects. These features are then associated in a learning process with certain words, resulting in an association of colour features with colour words, spatial features with prepositions, etc., and based on this, these words can be interpreted with reference to the scene currently presented to the video feed.

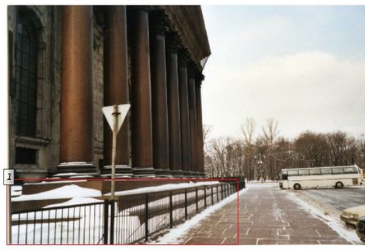
Of more recent work, that of Matuszek et al. (2012) is closely related to the approach we take. The task in this work is to compute (sets of) referents, given a (depth) image of a scene containing simple geometric shapes and a natural language expression. In keeping with the formal semantics tradition, a layer of logical form representation is assumed; it is not constructed via syntactic parsing

User: Draw a new card
System: *Ok.*

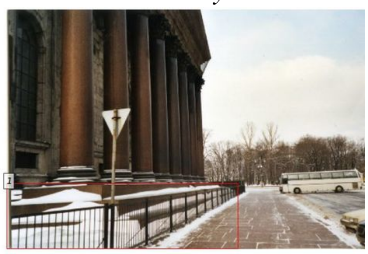


U: Find the rightmost pillar
S: *Let me see. Did you mean this?*

U: yes
S: *Ok.*
U: Find the pavement
S: *I'm going for this one. Correct?*



U: This is not a pavement
S: *Alright.*
U: 1 is a fence
S: *Ok.*
U: find the fence
S: *Let me see. Did you mean this?*



U: yes
S: *Thanks!*

Figure 1: **Example Interaction** *Reference candidates outlined in red and without label; selected candidates with numeric label. (Best viewed in colour.)*

rules, however, but by a learned mapping (*semantic parsing*). The non-logical constants of this representation then are interpreted by linking them to classifiers that work on perceptual features (representing shape and colour of objects). Interestingly, both mapping processes are trained jointly, and hence the links between classifiers and non-logical constants on the one hand, and non-logical constants and lexemes on the other are induced from data. In the work presented here, we take a simpler approach that directly links lexemes and perceptions, but does not yet learn the composition.

Most closely related on the formal side is recent work by Larsson (2015), which offers a very direct implementation of the 'words as classifiers' idea (couched in terms of type theory with records (TTR; (Cooper and Ginzburg, 2015)) and not model-theoretic semantics). In this approach, some lexical entries are enriched with classifiers that can judge, given a representation of an object, how applicable the term is to it. The paper also describes how these classifiers could be trained (or adapted) in interaction. The model is only specified theoretically, however, with hand-crafted classifiers for a small set of words, and not tested with real data. More generally, the claim that the ability to negotiate meaning is an important component of the competence of meaningful language use, which we also make here, has been forcefully argued for by Larsson and colleagues (Cooper and Larsson, 2009; Larsson, 2010; Fernández et al., 2011). (See also DeVault et al. (2006), who call this process *societal grounding* and outline a formal computational model of it.)

The second "dimension" in our semantic representations concerns language-to-language grounding. To explain within the framework of formal semantics how some statements can be necessarily true by virtue of meaning and not logical tautology (e.g., "bachelors are unmarried"), Carnap (1952) introduced *meaning postulates*, which are axioms that explicitly state connections between non-logical constants (e.g., $\forall x.bachelor(x) \rightarrow \neg married(x)$). The computational resource WORDNET (Fellbaum, 1998) can be seen as a large-scale realisation of this concept. It is a large database of word *senses*, different meanings that a word can have. Further semantic relations structure this lexicon (*antonymy, hyponomy, hypernymy, meronymy*). As described below, we use it as a starting point for encoding language-to-language grounding, together with the more directly perception-oriented feature norms of Silberer et al. (2013), which encode typical attributes ("is brown", "has feet") for about 500 concepts.

In the present work, our focus is on acquiring and using referential competence. On the ontological side, for now we simply use pre-compiled taxonomic/ontological resources. Methods exists for automating the construction of such resources (e.g., (Mitchell et al., 2015; Ganitkevitch et al., 2013)), some even using dialogue (Hixon et al., 2015). As another type of method, distributional semantics has recently become popular for the unsupervised acquisition of lexical relations (Turney and Pantel, 2010; Mikolov et al., 2013), particularly of the (typically rather vaguely specified) relation of 'similarity'. We will investigate the applicability of these methods in future work, but for now make use of the greater expressiveness and explicitness of more logic-inspired representations as used in WORDNET.

3 Overview of the Model

As stated in the introduction, a desideratum for the model is that it explains what kind of discriminatory power constrains meaningful language use, and how this power is acquired, defended and adapted in interaction. To make this more concrete (and to move from Putnam's tree example to a different biological kingdom), what we want to achieve is that our model can capture the knowledge required to deal satisfactorily both with (1-a) and (1-b).

(1) a. Find the Rottweiler in the picture.
    b. Peter walked past a Rottweiler. The dog was barking ferociously.

But what is this knowledge? For (1-a), this must be information connected with the visual appearance of the object that is to be identified; for (1-b), the required knowledge is that a Rottweiler is a type of dog, and hence that the definite noun phrase in the second sentence can refer to the object introduced into the discourse with the indefinite in the first. These types of knowledge can interact: We'd still be satisfied if, when presented with an image containing one Rottweiler and, say, five cats, the addresse points at the Rottweiler, even if they don't actually know what distinguishes Rottweilers from other breeds of dog and all they knew was what visually distinguishes dogs from cats.

We take the basic idea from Marconi (1997) that there is a categorical difference between these types of knowledge. Marconi (1997) labels these aspects of lexical competence *referential* and *inferential*. While our focus in the work presented here is also on reference, we would argue that the distinction is more generally one between *know-how* and *know-that*, with the former covering the knowledge involved in executing actions ("cycling", "drawing an elephant") as well, and we will refer to the types with these labels. These "two dimensional" lexical semantic representations then must be composed into representations of phrases, where the composition process as well as what went into it must be open to justification and critique in interaction. We address these parts of the model in turn.

4 Two-Dimensional Lexical Semantics

(2) sketches the lexical entry for 'Rottweiler' with its two basic components, "know-how/referential" and "know-that/ontological", as it will be explained in the following.

(2) $\begin{bmatrix} \texttt{Rottweiler} \\ \textit{kh/ref} : \lambda\mathbf{x}.f_{rt}(\mathbf{x}) \\ \textit{kt/ont} : wn.hyponym, wn.hypernym, etc. \end{bmatrix}$

4.1 Visual/Referential know-how

We follow Kennington and Schlangen (2015) and represent (and learn) visual-referential knowledge as classifiers on perceptual input. We briefly review their model here.

Let w be a word whose meaning is to be modelled, and let $\mathbf{x}$ be a representation of an object in terms of its visual features. The core ingredient then is a classifier that takes this representation and returns a score $f_w(\mathbf{x})$, indicating the "appropriateness" of the word for denoting the object. In (Kennington and Schlangen, 2015) and below, the classifier is a binary logistic regression and the score can be interpreted as a probability. Training of the classifier will be explained below.

Noting a (loose) correspondence to Montague's (1974) intensional semantics, where the intension of a word is a function from possible worlds to extensions (Gamut, 1991), the *intensional* meaning of w is then defined as the classifier itself, a function from a representation of an object to an "appropriateness score":[1]

$$[\![w]\!]_{obj} = \lambda\mathbf{x}.f_w(\mathbf{x}) \quad (1)$$

(Where $[\![.]\!]$ is a function returning the meaning of its argument, and $\mathbf{x}$ is of the type of feature given by f_{obj}, the function computing a feature representation for a given object.)

The *extension* of a word in a given (here, visual) discourse universe W can then be modelled as a probability distribution ranging over all candidate objects in the given domain, resulting from the application of the word intension to each object ($\mathbf{x}_i$ is the feature vector for object i, $normalize()$ vectorized normalisation, and I a random variable ranging over the k candidates):

$$[\![w]\!]^W_{obj} = normalize(([\![w]\!]_{obj}(\mathbf{x}_1), \dots, [\![w]\!]_{obj}(\mathbf{x}_k))) = normalize((f_w(\mathbf{x}_1), \dots, f_w(\mathbf{x}_k))) = P(I|w) \quad (2)$$

4.2 Taxonomic/Ontological know-that

As mentioned above, for now we use pre-existing resources as source of the initial ontological knowledge. There is some selection of available sources besides WORDNET (e.g., Freebase (Bollacker et al., 2008) and ConceptNet[2]), but we start with the former, as it is well-curated and stable. It provides us mostly with hypernomy (or "*is a*") relations. Notoriously, these can contain rather arcane categories; (3) shows this information for the lexical entry for "Rottweiler" with the less common categories (such as *placental* or *chordate*) left out.

(3) $\begin{bmatrix} \texttt{Rottweiler} \\ \textit{kt/ont/hyp} : \text{shepherd\_dog|working\_dog|dog|...} \end{bmatrix}$

An additional, but with 509 entries compared to the over 200k entries of WORDNET much smaller information resource is the set of feature norms of McRae et al. (2005), a collection of attributes typically associated with a given object. (We use the version prepared by Silberer et al. (2013), which is filtered for being backed up with visual evidence.)

This resource does not contain an entry for *Rottweiler*, but one for *dog*, which is shown in (4).

[1](Larsson, 2015) develops this intension/extension distinction in more detail for his formalisation.

[2]`http://conceptnet5.media.mit.edu`

$$
(4)\quad \left[\begin{array}{l}
\texttt{dog} \\
\textit{kt/ont/isa} : \text{animal}|\text{mammal} \\
\textit{kt/ont/properties} : \\
\left[\begin{array}{l}
\textit{anatomy/has} : \text{mouth, head, whiskers,} \\
\text{claws, jaws, neck, snout, tail, 4\_legs, teeth,} \\
\text{eyes, nose, fur, ears, paws, feet, tongue} \\
\textit{behaviour} : \text{walks, runs, eats} \\
\textit{colour\_patterns} : \text{grey, black, brown, white} \\
\textit{diet} : \text{drinks\_water}
\end{array}\right]
\end{array}\right]
$$

We have explored two other kinds of automatically acquired lexical relations, but postpone their description until we have described the data sets that we used for our implementation.

5 Composition

5.1 Visual/Referential know-how

In the Kennington and Schlangen (2015) approach, composition of visual word meanings into phrase meanings is governed by rules that are tied to syntactic constructions. In the following, we only use simple multiplicative composition for nominal constructions:

$$[\![[_{nom} w_1, \dots, w_k]]\!]^W = [\![\text{NOM}]\!]^W [\![w_1, \dots, w_k]\!]^W = \circ_{/N} \left([\![w_1]\!]^W, \dots, [\![w_k]\!]^W\right) \quad (3)$$

where $\circ_{/N}$ is defined as

$$\circ_{/N} \left([\![w_1]\!]^W, \dots, [\![w_k]\!]^W\right) = P_\circ(I|w_1, \dots, w_k)$$
$$\text{with } P_\circ(I = i|w_1, \dots, w_k) =$$
$$\frac{1}{Z}(P(I = i|w_1) * \cdots * P(I = i|w_k)) \text{ for } i \in I \quad (4)$$

(Z takes care that the result is normalized over all candidate objects.)

To arrive at the desired extension of a full referring expression—an individual object, in our case—, one additional element is needed, and this is contributed by the determiner. For uniquely referring expressions ("the red cross"), what is required is to pick the most likely candidate from the distribution:

$$[\![the]\!] = \lambda x. \operatorname*{arg\,max}_{Dom(x)} x \quad (5)$$

$$[\![[the]\ [_{nom} w_1, \dots, w_k]]\!]^W = \operatorname*{arg\,max}_{i \in W} \left[\ [\![[_{nom} w_1, \dots, w_k]]\!]^W\ \right] \quad (6)$$

5.2 Taxonomic/Ontological know-that

Composition of the ontological information is less fully developed at the moment. We can describe the requirements, though. For a phrase like "*the black dog*", we would want the general terminological knowledge encoded in (4) ("a dog is an animal, and (typically) is grey or brown or . . . ") to be specialised to this particular instance ("this dog is an animal . . . ") and the disjunctive attribute information to be restricted ("... and it is black"). This corresponds to the distinction between 'terminological axioms' in the so-called TBox and 'assertional axioms' in the ABox in Description Logic (Krötzsch et al., 2014), which should also have the necessary expressiveness to realise this composition process.

6 Interaction

The final component is the actual meta-linguistic interaction that takes as topic the adequacy of the predictions made by the other components. As, unlike in distributional semantics or in approaches to language/image matching using deep learning approaches (e.g., (Hu et al., 2016; Mao et al., 2016)), we specify the composition process explicitly, we have access to all its intermediate steps. We can hence provide justifications for object selection decisions that can adress the individual words as well as their composition. This will be described in more detail in the next section.

7 Implementation

7.1 Learning Visual Meanings

The visual classifiers are trained on large sets of images that are segmented into objects, for which referring expressions exist. This is described in more detail for a static recognition task in (Schlangen et al., 2016). We outline the process here, as the trained models form the basis for the interaction, which is the contribution of this paper.

One dataset is the SAIAPR/ReferIt set. It contains of 20k images with a tourism theme (Grubinger et al., 2006) for which object segmentations (Escalante et al., 2010) and, for these objects, referring expressions are available (120k altogether; Kazemzadeh et al. (2014)) . The second dataset is based on the "Microsoft Common Objects in Context" collection (Lin et al., 2014), which contains over 300k images with object segmentations (of objects from 80 pre-specified categories), object labels, and image captions. This has also been augmented with referring expressions by the same group as (Kazemzadeh et al., 2014), in as yet unpublished work (the authors provided us with this data on request). For evaluation, we downloaded 20k images from those synsets covered in ImageNet (Deng et al., 2009) for which object segmentations were available in ImageNet.

As representations of the images we used the final fully-connected layer before the classification layer of a deep convolutional neural network, "GoogLeNet" (Szegedy et al., 2015), that was trained on data from the Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) from the ImageNet corpus (Deng et al., 2009) to extract features, to give us a 1024 dimensional representation of the region. We augment this with 7 features that encode information about the region relative to the image: the (relative) coordinates of two corners, its (relative) area, distance to the center, and orientation of the image. The full representation hence is a vector of 1031 features.

In the training process then, each word occurring in the training part of the referring expression collection is paired with the region it describes (and with 5 randomly selected regions for which this word never ocurrs in the description) to train a logistic regression classifier. Here, we use the classifiers trained on the combination of the two datasets, which yields models for 933 words. As shown in (Schlangen et al., 2016), this model reaches accuracy on a static visual reference resolution task that is at the state of the art of end-to-end deep learning models. Our focus here, however, is on the use of these models in interaction, which we turn to next.

7.2 Interacting with Semantics

To get an impression of whether interaction ("meaning negotiation") can improve not only performance of the system but also the impression of considered language use, we implemented a simple dialogue system that uses these models. We interfaced it with a popular chat tool,[3] so that interactions can be done remotely.[4]

Together with a simple dialogue logic that handles misunderstandings and state updates (Larsson and Traum, 2000), we implemented the following commands (each accessible through a variety of formulations, only one of which we give here):

- *draw a new card*, which presents a new image with a natural scene to the user, with the candidate objects (for which manual segmentations are available) outlined.

- *draw a new collage*, which presents a collage consisting of 6 sub-images that each contain a single object; the understanding here is that

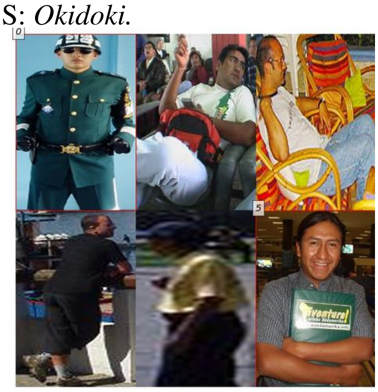
[3]Slack, `http://www.slack.com`

[4]By the time of the conference, we will release all code as open-source. We will also provide a publicly accessible installation of the system and will use it to collect more labels from users, and to learn online from them.

User: get a new collage men
System: *Ok.*

U: Find the man with the book
S: *Let me see. Did you mean this?*

U: yes
S: *Thanks!*
U: explain
S: *Okidoki.*



S: *0 is best for "man"*
S: *5 is best for "book"*
S: *Overal best: 5*
S: *Rank of region 5 for "man": 4*
S: *Rank of region 5 for "book": 1*

Figure 2: **Example Interaction II** *Reference candidates outlined in red and without label; selected candidates with numeric label.*

User: what do you associate with trees?
System: *palm, church, tree, house, fountain, beach, grass, statue, road*
U: what looks similar to trees?
S: *tree, plants, leaves, grassy, bushes*
U: what looks similar to bicycle?
S: *bike, rider, four, biker*

Figure 3: **Example Interaction III**

the system "sees" the same sub-images, but not necessarily in the same layout.

- *find X* (e.g., "find the person in the background on the left"), to which the system responds by outlining its selection and asking whether this selection is correct.
- *explain*, which asks for explanation of the selection decision, which the system provides by listing what fits best to each word individually, and then how the selection is composed.
- *what looks similar to X?*; this gives indirect information about the visual classification that the system has learned, as described below.
- *what do you associate with X?*, which similarly gives information about the learning situation, as also described below.
- *X is (not) Y* (e.g., "this is not a boat", or "3 is not black"), which adds this information to the set of labels, which can then be used for adapting the classifiers.

Information about what looks similar is computed as follows: We randomly select 2000 image regions from a held out set and run all word classifiers on them. This results in 2000 responses (probabilities of fit) for each word, or in other words a 2000-dimensional vector that represents the reactions of this word-classifier to the sample objects. Similarity can then be computed in the usual way as a relation between vectors (we use the cosine); but the resulting type of similarity is a visual one. (More details and evaluations will be given elsewhere.)

The associative information is compiled by computing pointwise mutual information between words ocurring in descriptions of objects within the same scene. This brings objects that often occur together in the same image (such as houses and roads) together.

So far, we have run informal tests during development of the system. In one such test with a naive user, the user interacted for 30 minutes and added more than 40 facts in this time. In a post-experiment questionnaire, they ranked the system highly for the interest that the interactions generated, and they indicated that the interaction helped them form hypotheses about the word meanings learned by the system, better than looking at examples of successful and unsuccessful reference resolutions would have. More formal and comprehensive testing is of course still required.

8 Conclusions

We have outlined a model of grounded semantics that combines perceptual grounding with ontological grounding. This model serves as the basis of a dialogue system that can play a simple reference game, and can provide justifications for the decisions it makes, and accept corrections.

The visual-perceptual part of the model is fairly well-developed, and has been shown elsewhere to achieve good accuracy on an offline task (Schlangen et al., 2016), and has shown some promise as a bidirectional model that can also be used for generation (Zarrieß and Schlangen, 2016). Based on the preliminary tests reported here, embedding it in an interaction seems promising. Much still remains to be done, however. First, the way how what we call the lexical 'know-how' here and the 'know-that' is combined needs to be more fully formalised, and the reasoning this requires and enables must be described. Second, the taxonomic and ontological knowledge should also be acquired in interaction and be negotiable in interaction. The implementation should form a good basis for making these extensions.

Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Thanks are due to the anonymous reviewers for their helpful comments.

The image processing code can be found at `https://github.com/dsg-bielefeld/image_wac`.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of ACM SIGMOD*.

Rudolf Carnap. 1952. Meaning postulates. *Philosophical Studies*, 3:65–73.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantic Theory 2nd edition*. Wiley-Blackwell.

Robin Cooper and Staffan Larsson. 2009. Compositional and ontological semantics in learning from corrective feedback and explicit definition. In *Proceedings of "Diaholmia" (semdial 2009)*, pages 10–14.

Jia Deng, W. Dong, Richard Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, USA, July.

Hugo Jair Escalante, Carlos a. Hernández, Jesus a. Gonzalez, a. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT 2011)*, Barcelona, Spain, July.

L. T. F. Gamut. 1991. *Logic, Language and Meaning: Intensional Logic and Logical Grammar*, volume 2. Chicago University Press, Chicago.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDN: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 13–23, Genoa, Italy.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *NAACL 2015*.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of CVPR 2016*, Las Vegas, USA, June.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Beijing, China, July. Association for Computational Linguistics.

Markus Krötzsch, František Simančík, and Ian Horrocks. 2014. A description logic primer. In Jens Lehmann and Johanna Völker, editors, *Perspectives on Ontology Learning*, chapter 1. IOS Press.

Staffan Larsson and David Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, pages 323–340.

Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 83–90, Poznan, Poland.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of logic and computation*, 25(2):335–369.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C.Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR 2016*, Las Vegas, USA, June.

Diego Marconi. 1997. *Lexical Competence*. MIT Press, Cambride, Mass., USA.

Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *ICML 2012*.

Ken McRae, George S. Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments & Computers*, 37(4):547—-559, feb.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pages 1–9.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Hilary Putnam. 1973. Meaning and reference. *Journal of Philosophy*, 70:699–711.

Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. 2002. A trainable spoken language understanding system for visual object selection. In *Proceedings of the International Conference on Speech and Language Processing 2002 (ICSLP 2002)*, Colorado, USA.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3).

Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciene*, 9(8):389–396.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*, Berlin, Germany, August.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *ACL 2013*, pages 572—-582, Sofia, Bulgaria.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Richmond H. Thomason, editor. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven and London.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of ACL 2016*, Berlin, Germany, August.