An Outline of a Model of Situated Discourse Representation and Processing

David Schlangen
Computational Linguistics // Department of Linguistics
University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Situated discourse provides opportunities for both exophoric and endophoric reference, hence requiring mechanisms linking both language and world, and language and (representation of) language. It is also, arguably, the primary site for learning about these mechanisms, which I take to be the meaning of words. I outline a model that brings together a treatment of discourse reference and a learning-based treatment of exophoric reference, based on the metaphor of "mental files". It combines a (rudimentary) model of visual attention with a model of exophoric reference resolution, built on a pluralistic model of word meaning that accounts for learning from ostension, learning from overhearing and learning from definition. Lastly, it offers several pathways to making inferences, accounting for fast judgements and slow justifications. At this stage, the model mostly provides a way of structuring the task of processing situated discourse, meant to be extensible both in the direction of stricter formalisation as well as in the direction of further implementation, the beginnings of which are reported here.

1 Introduction

Let's start with a story.

> Ann and Bert are taking a stroll in the park. *"Look at the dog, over at the waterfountain"*$_{(1)}$ Ann says, looking at a man carrying a dog in his arms. *"I know, isn't it cute!"*$_{(2)}$, Bert replies.
>
> A short while later they run into Chris and her young son, Dale. *"We just saw a man carrying a dog,"*$_{(3)}$ Bert exclaims. *"The cutest poodle ever!"*$_{(4)}$, adds Ann.
>
> *"Actually, I don't think that was a poodle."*$_{(5)}$, says Bert, who can't stop himself. *"It was too tall. I think it was a labradoodle."*$_{(6)}$ Ann knows that Bert is a serious hobby cynologist, so she says *"Oh. I guess you're right. You're the expert here."*$_{(7)}$
>
> After exchanging more pleasantries, they go their separate ways again, Chris and Dale walking into the direction from which Ann and Bert just came. And sure enough, shortly thereafter they see a man who is holding a beautiful white-ish dog on a leash. Chris happens to know the man, who had previously told her about his new dog, and so she thinks to herself: *"They were talking about Fredo!"*$_{(8)}$ Dale, however, thinks *"So that's what a labradoodle looks like."*$_{(9)}$

This is admittedly not a particularly exciting story. But mundane as it is, no existing computer system can model all the behaviours exhibited here, and neither can any formal model of situated discourse. So, what is it that Ann, Bert, Chris, and Dale need to be able to do to play their respective parts in this little story, and how could that be modelled in a machine?

- For exchange (1)/(2) to work (referring to individual utterances from the story via their subscripted numbers), both Ann and Bert need to be able to parse their environment into objects, and to recognise one of these as being appropriately described by *dog*. Furthermore, after (1), that dog can also be referred to by *it*, which wasn't the case before.
- To understand (3)/(4), Chris and Dale need to be able to connect *poodle* and *dog* in such a way as to let the phrases in which they occur refer to the same entity, or at least recognise that to do so is Ann's intention with (4). Unlike for Ann and Bert in (1)/(2), for Chris and Dale that entity is not also perceptually available. For Ann and Bert, it is not perceptually available *anymore*, but they will have a visual memory of it.
- (5)/(6)/(7) shows that categorisation decisions are *negotiable*: Bert denies applicability of *poodle* to the entity from discourses (3)/(4) and (1)/(2), provides reasons for doing so, and suggests a better category. (7) indicates that expertship is a reason

for accepting a proposed revision. We can assume that an effect of this interaction is that Ann changes her understanding of *poodle*, and possibly that of *labradoodle* as well, and perhaps Chris and Dale do so as well with their respective ones.

• (8) and (9) finally show some later consequences of the interaction. (8) shows that Chris can now make a connection she couldn't make earlier (the dog that Ann and Bert saw is Fredo, the new dog of her acquaintance). (9) shows Dale attaching visual information (from perceptual acquaintance) to a concept that previously, we assume, had only been introduced verbally to him.

We can derive from this analysis some desiderata for a model of agents that can participate in such situated discourse:

A) It needs to provide visually present real world objects as possible referents, as well as objects only introduced via discourse, as well as entities that previously were introduced either way.

B) It needs to provide word meanings that guide the mechanisms by which utterances are linked to either real world objects ("dog" in (1)), or to discourse referents ("dog" and "poodle", in (3)/(4)); which moreover are accessible for discussion and revision (5)/(6)/(7).

C) It needs to provide a way in which word meanings are composed into utterance meanings, and a way in which those are composed into discourse meanings.

In this paper, I attempt to outline a model that meets these desiderata. A particular focus in doing so is on the potential for recruiting current computational work for specifying the *mechanisms* that would be needed in a realisation of a dynamic model of situated discourse, and on arguing for a particular way of factorising the problem. This comes at the cost of a lack of formal rigour (for example, a lack of a model theory that would precisely specify the *conditions* under which an agent would evaluate a statement as true). This is something that I hope can be delivered later, where appropriate.[1]

2 Overview

Following closely the structure of the desiderata from the previous section, the proposal consists of several interlocking parts. The job of the **context representation component** is to provide the entities that the interpretation process links together. To perform this linking, it is not enough to just have access to (mental representations of) entities, there also needs to be further information connected to them. To describe this structured keeping of information, I will make use of the *file metaphor*, according to which each entity is represented by one file, on which additional information can be "written".[2] To give an illustration, Figure 1 shows what Bert's representation of the situation after utterance (1) looks like. A *file card* (along the lines of Heim's (1983) file change semantics) has been created for the discourse entity that (1) introduced; as part of the situated interpretation, Bert has been able to link that to one of the *object files* that his perceptual system created to represent the visual scene. Had he recognised the object as a previously know entity, he could also have made a link to what we call an *entity / event (e/e) file*; but here this is not the case.

What enabled Bert to make the connection are relevant *concept files*, which are part of the **conceptual component** of the model. Our proposal here is to account for the relevant desiderata rather directly, by representing the conceptual content that enables categorisation of perceptual objects (in the example here, identifying one object as a dog, and as white) and that which enables relating concepts (in the story, *dog* and *poodle*), separately. Moreover, we assume that at least parts of this content are accessible to their owner, and can be discussed and voluntarily revised (as in (5)-(7) of the story). The concept files come categorised into domains of knowledge, and this can factor in the decision to defer to a judgement (and revise one's own concept), or not.

Of course, it's not single words that trigger the creation of links in the context representation, it's the composition of words into phrases and utterances. The final part of the model then is the **composition component**, which composes utterance meanings out of the recognised structure and the activated concept files.

The remaining sections describe these components further and speculate on the computational mechanisms that might be used to create and maintain these representations, and discuss where cur-

[1] Maybe what I'm proposing here is not so much an actual model than more a research program, or a proposal for how to put together some existing smaller parts into one larger whole.

[2] This metaphor has a long, and wildly branching, history in linguistics (e.g., (Karttunen, 1969; Heim, 1983)) and philosophy (see the recent overview in (Recanati, 2012)), to the extent that not any single precursor can be declared here.

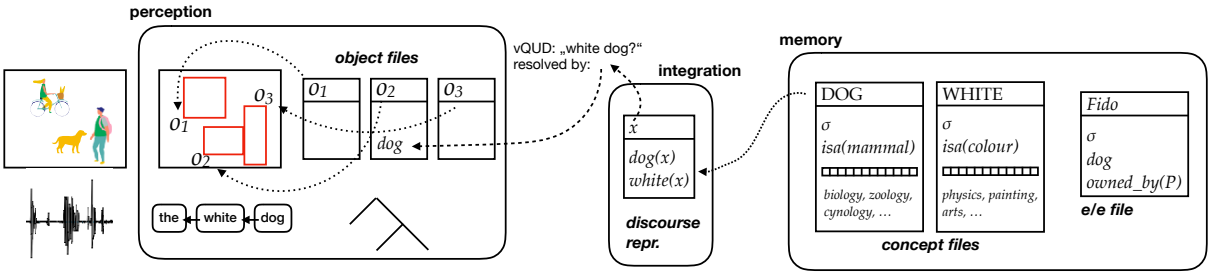

Figure 1: Overview of the Representational Components of the Model

rent implementations fall short.

3 Context Representation

As mentioned above, this is the component that is responsible for providing the mental representations that are linked together in the process of discourse interpretation. At the level of elaboration achieved here, I am assuming that any standard model of discourse representation can handle the part of representing discourse referents. The file metaphor I am using throughout here matches more closely with Heim's (1983) model, but the graphic representation in Figure 1 is more clearly inspired by DRT (Kamp and Reyle, 1993). (And a more sophisticated model like SDRT (Asher and Lascarides, 2003) would be needed to account for all the discourse coherence phenomena that I gloss over here.)

What is less often spelled out is how the visual scene can be made accessible in discourse representation (but see the related work discussed below). In the proposed model, the visual scene is represented via what I call *object files*. How do these object files connect to the actual objects in the world? I assume here that the indices with which the object files are indexed are created automatically by the perceptual system, and that these indices are pre-conceptual in that when created they do not represent more than raw objecthood.[3] I take this from Pylyshyn (1989), where a similar concept was introduced to account for the experimentally tested ability of human subjects to track a (limited, but not small) number of objects moving across a screen. Interestingly, this tracking worked even if some features of the objects (such as colour, shape, and texture) changed continuously; in fact, these changes were often not even noticed. Pylyshyn (1989) calls these tracking representations FINSTs ("Fingers of Instantiation"), with the idea that they function as indices for visually presented objects, individuating them across time and providing spatial information about them, without anything more needing to be recognised and represented about the objects. In later work, Fodor and Pylyshyn (2015) recognised the utility of this conception for providing a basis for conceptual reference, and it is this use that I assume here as well.

The indices or FINSTs only represent the presence of presence of *some* kind of object. Clearly, in many situations, this it is not enough for an agent, and it is of interested to know *which kind* exactly it is. However, any given object will be a member of many kinds. I make the assumption that the kinds that currently matter are determined by the current interests of the agent. As a simple device for specifying these interests, I introduce into the model the notion of a "visual Question under Discussion" (vQUD), in analogy to the discourse structuring device "QUD" proposed by Ginzburg (1995, 2012). This device is to act as an interface between bottom-up object *recognition* and top-down object *classification*.

In many situations, this question may default to something like "what's this?", which can perhaps be resolved by classification into what (Rosch, 1978) called *basic level categories*. But the mechanism should also account for how differently the world is parsed if there are pressing current concerns: If approached by a hungry hyena, you need a fast answer to "what can I throw?"; when hungry yourself and in the absence of any wild beasts, the salient question is a quite different one. To go back to our example from the introduction, and also to bridge to the next section, the vQUD mechanism shall also serve as the interface to reference resolution, where the question is directly provided by the

[3] The notion of "objecthood" that I am appealing to here is compatible with that investigated by Spelke (1994); Spelke and Kinzler (2007); Carey (2009) as likely to be innate, i.e. roughly "discrete, bounded entity that moves coherently, if force is exerted or produced by it".

task of utterance context integration.[4]

In any case, resolving the question on vQUD requires the application of relevant concepts to the objects. The objects are accessible via their indices / FINSTs, and the results of the classification are collected in the aforementioned *object files* that are connected to them.[5] These are meant as a representation of the current conceptual information about the object, active for as long as the current scene is attended to. (Let's assume that information about the perception event can be transfered into long-term memory in the form of an "event / entity file", as mentioned above. Also, re-recognition of an individual can be achieved by linking such an e/e file to an active object file; but more on that later.)

Lastly, let us also assume that an agent typically optimistically supposes that objects they recognise are also recognised by co-present agents.

Implementation With the dramatic improvements in quality that computer vision models have seen in recent years, it might seem that an implementation of this part of the model could be taken directly off the shelf. And it is true that current models like Faster R-CNN (Ren et al., 2017) and YOLO (Redmon et al., 2016) achieve high accuracy in segmenting and labelling still images. However, these models work by overgenerating proposed object regions and subsequent filtering through categorisation from a fixed set of categories. This misses two properties of the model as described above: a) it neglects that are always many valid classification possiblities, and which one is to be selected depends on current purposes (the reason why we've introduced the vQUD mechanism), and b) it restricts the set of classes to a prespecified finite set, whereas we want it to be possible to incrementally learn new classes. Hence, some adaptations to the basics of these algorithms would be required in an implementation of the model.

In the computational experiments we have run so far, we have skipped over this step and start with the segmentations provided by typical vision corpora (e.g., COCO, (Lin et al., 2014)), where objects are marked by bounding boxes. We take these, and the identifiers they come with, as simulating the output of the initial perception stage, providing the indices to which the object files are linked.

4 Word Meanings and Concept Files

As mentioned in the introduction, the concepts files collect the information that is required to create the links between representational objects; that is, between discourse referents and object files, discourse referents and antecedents, and discourse referents and e/e files. I first discuss some of the desiderata of what mental concepts should address, before presenting the proposal.

The Challenges I follow Bloom (2000) in assuming that learning the meaning of a word entails associating a form with a concept. If we understand concepts to be mental representations, as I am doing here, there is a problem. Mental representations are private, but words are public entities belonging to a public language, and so we will need a way to make the *content* of concepts publicly negotiable. Secondly, we need these concepts to do different things—or at least I will analyse these as being different—namely to support both exophoric and endophoric reference, of the kinds illustrated with our introduction story. (More generally, we will want them to support the computation of denotations and of material inferences.) Connected to this is that we want to capture different ways of learning—or at least we will analyse these as different—namely learning from observation of successful reference, learning from explicit definition (through linguistic explanation), and learning from implicit definition (through linguistic context). Lastly, we need to set them up in such a way that they are individuated in the right way, and the challenge to answer here is what Perry (2012) calls the 'co-reference and no-reference' problems (which are, of course, Frege's (1892) puzzles): How can an agent who possesses the words/concepts *Hesperus* and *Phosphorus* potentially learn something new when told that those name the same object? And what makes the concepts of Santa Claus and of unicorns different, given that they both refer to

[4] Below, as in this paragraph, I represent the questions on vQUD using natural language, without intending to commit to assuming a language of thought. But the relation between the kinds of categorial decisions one can make privately and those that one can make publicly via language, however, is of course a fascinating subject, left here to be explored in future work.

[5] Borrowing this label from Kahneman et al. (1992), who give a vivid example of why separating object *individuation* and object *classification* seems to fit the phenomenology of perception: "Imagine watching a strange man approaching down the street. As he reaches you and stops to greet you he suddenly becomes recognizable as a familiar friend whom you had not expected to meet in this context." It is the classification that changes here, not the identity of the object (as object) that you had tracked all along.

nothing?[6]

The Proposal The proposal I make here is to reify the observed (or claimed) differences, as it were. Namely, I assume that these different facets are indeed accounted for by different representations, brought together in what I call *concept files*.[7] Figure 1 gave the example of Bert's concept file for DOG (repeated below as Figure 2). The σ stands for the first component, the knowledge that allows an agent to recognise an object as falling under the concept or not. I follow the 'words as classifiers' model here (Schlangen et al., 2016; Kennington and Schlangen, 2015) (see also (Larsson, 2015, 2011)) and realise this technically through statistical classifiers that generalise via supervised learning and error-driven adaptations. (See below for technical details.) Such knowledge, I assume, can be picked up whenever successful reference is observed. This can be, but doesn't alway have to be, in episodes of ostensive teaching; I only assume that the learning agent must be capable of understanding what the intended referent was (so Bloom's (2000) arguments against associationism do not bite here).

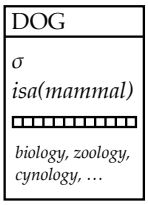

Figure 2: A Concept File for DOG

The second component is what accounts for the discussion in (5) and (6) of our introductory example, where Bert provides factual information about the concepts POODLE and LABRADOODLE. I assume that such factual information is indeed encoded as such, and can be recalled when needed, for example to provide justifications for classification decisions. (As I argued before in (Schlangen, 2016).) This kind of knowledge weaves concepts into a network,[8] and provides the kind of information that can support resolving discourse anaphoric phenomena like the intended co-reference between "the poodle" and its antecedent "a dog" in (4).[9]

I also assume that connections between concepts are learned and represented in less explicit ways, namely from distributional information that gets reduced into vector representations of word meanings. (This is what the row of boxes in the concept files in Figure 1 is meant to represent.) I am swayed here not only by the psycholinguistic evidence that such distributional information figures in meaning representations (see, e.g. (Andrews et al., 2009; McDonald and Ramscar, 2001)), but also use this to account for the phenomenological difference between trying to produce a definition of a concept and making a fast judgement of similarity between concepts. (More on that below in Section 5.)

Finally, and connected to the second component described above, I assume that it is part of conceptual knowledge to have a rough idea about the domain or field of knowledge that a concept belongs to. This can then serve as a guide for disputes about word meanings (or "litigations" of word meaning, (Ludlow, 2014)) on how to weigh other people's opinions, with the limiting case being to defer to the expert (Keil and Kominsky, 2015).

Challenges Addressed? First, the desideratum of accounting for exophoric and endophoric reference. Here, in the basic cases, the responsibilities are divided: Exophoric reference is resolved via the classifiers that work on perceptual input, and the resolution of endophoric reference, insofar as it needs inference, is supported by the semantic or distributional knowledge in the concept file. I do however assume that cross-overs are possible: For example, in fine-grained visual classification (Dubey et al., 2018), recognition can happen via explicitly recalling distinguishing features (e.g., between bird species) and recognising those, deriving from this the visual classification decision.

This carries over to the learning of the components of the concepts files. While each component primarily seems to be connected with one type of learning situation described above (observing a reference intention with learning a classifier; receiving semantic information with learning semantic information; observing linguistic contexts with learning a continuous representation), initial experiments (see below) make us hopeful that cross-modal /

[6] This list is only the beginnings of the trouble with concepts, as (Murphy, 2002; Margolis and Laurence, 2015) helpfully review, but it shall be enough for this first start.

[7] Marconi (1997) provides a book-length argument for a similar separation which I took as an inspiration, but develop in a different technical direction here.

[8] Whereupon the file metaphor starts to falter a bit; but think of relational databases and their links between records.

[9] I leave possible connections to the "theory theory" model of concepts (Morton, 1980) on the one hand, and Donald Davidson's (1986) coherentism unexplored for now.

zero-shot learning can be realised. For example, visual recognition on first encounters of instances of otherwise familiar concepts ("book knowledge") can back off to recognition via properties stored in semantic knowledge; semantic knowledge can be induced from perceptual similarities; distributional knowledge can support visual classification, and be discretised into semantic knowledge.

Allowing these types of knowledge to be incrementally adapted, through interactions such as in the example from the introduction, opens up the private concept to public inspection and debate. This then connects the individual's concept with the public (but virtual) intension. Of this, at least in certain domains, experts can be the custodians. So, while concepts are in the head(s), *meanings* (understood as the public norms governing use), in this model "ain't" (Putnam, 1975).

Then, there are Hesperus and Phosphorus.[10] What happens when an agent, who previously didn't know it, learns that these are names for the same entitiy? In this model, suprisingly little: the agent just has to bring together the respective files. That two files have to be merged ("Hesperus is Phosphorus"$_{(8)}$) is valuable information, whereas touting the identity of a file with itself ("Hesperus is Hesperus"$_{(9)}$) is not. Similarly, accounting for attitude reports is straightforward: They index the file that the agent calls on, and the label they use to access it, and so substitution of a co-extensive term (e.g., turning "A believes (9)" into "A believes (8)") describes a different mental state.

Finally, *no-reference*. In this model, we can assume that concepts of this type (like UNICORN) have as part the knowledge that no instances of them will ever be encountered. Nevertheless, *pretend* semantic information, and real distributional information, can still be encoded, so that the concept can be used meaningfully.

But with all this, aren't concepts too fine-grainedly individuated? It looks like there now isn't just one type of sense, but many? Can two individuals ever have the same conceptualisation connected to the same word? (They can't have the same *concept*, as concepts are mental entities.) Here, I am inclined towards a pragmatist reply: Whether two individuals connect sufficiently similar concepts with the same word shows in the consequences of their having these concepts. And this is again where dialogue comes in, and learning. If the agents encounter a situation where they appear to be disagreeing on the applicability of a term, they can discuss, and, hopefully, reach an agreement that makes one, or both, of them adapt their concept. Where Fodor and Pylyshyn (2015) seem to assume that any single experiential difference between two agents must make their concepts incommensurable (if the inferential roles of the concepts are what individuates them), it seems to me unlikely that the learning methods sketched here, run on what is fundamentally very similar data (if both agents live in the same language community), even if on the instance level fully disjunct, yield dramatically different results that cannot be made commensurable in clarification dialogues of the type illustrated in the introduction. But this is an empirical question that a large-scale implementation could begin to test.

Implementation The components of the concept files individually are well-established in computational work. I have already mentioned some work on the words-as-classifiers model (Schlangen et al., 2016; Kennington and Schlangen, 2015) (see also Matuszek et al. (2012) for a related approach). Representing conceptual information in so-called *knowledge graphs* is a thriving field (see (Hogan et al., 2020) for a recent review); computing continuous representations of words from distributional information, following Landauer and Dumais (1997); Mikolov et al. (2013), perhaps even more so. (These last two fields, however, do not seem to be concerned much with incremental learning, which I take to be an indispensable part of any realistic model of situated interaction.)

Zero-shot learning, as it is called in the machine learning community, where information from one modality (e.g., text) is utilised for a task in another (e.g., visual classificaiton) is also a field that has relevance for this model. An approach like that of Lampert et al. (2014), where a visual classifier is constructed out of semantic knowledge directly offers a connection between these two types of knowledge in our concept files. Zarrieß and Schlangen (2017b) have explored the use of distributional information in the words-as-classifiers model, whereas Zarrieß and Schlangen (2017a) have shown that referentially structured contexts can yield improved continuous representations. A

[10] These, of course, are *names*, which in the model sketched in Figure 1 would be the labels of what we called *event / entity files*. For the purposes of this discussion, however, we can view them as concepts that mandatorily have a singleton extension.

thorough analysis of the structure of meaning negotiation interactions is still needed, but some pioneering (formal) work is being done (Larsson and Myrendal, 2017). Previously, I presented a proof-of-concept implementation that makes use of visual classifiers and semantic knowledge in simple justification interactions (Schlangen, 2016).

To sum up, parts, and partial connections, between the proposed components, have been explored in computational work. What is still missing is an attempt to fully combine these efforts along the lines sketched here.

5 Composition into Utterance Meanings

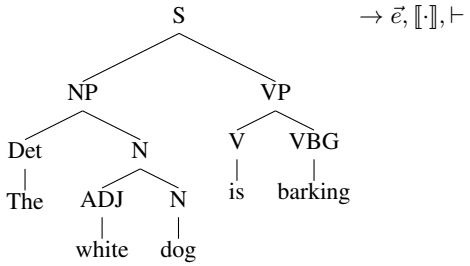

Figure 3: A single parse supporting three representations / uses

In keeping with the representational pluralism (or promiscuity?) of the previous section, I assume that there are several ways in which representations for larger expressions are computed, and used. The common basis for all of them is a syntactic analysis of the expression,[11] as in Figure 3.

First, I assume that a logical form is computed (e.g., (Sag and Wasow, 1999)) and integrated into the discourse representation (e.g., (Kamp and Reyle, 1993)), in the usual way. Besides other things, this representation forms the input to a within-context verification mechanism, which, for the visual context, is expressed via the vQUD as described above. For this, the interpretation functions for the non-logical constants are provided by the classifiers in the concept files. Following Schlangen et al. (2016), I assume that the contributions of the individual concepts are composed into a response for the whole phrase, on the basis of which (probabilistic estimates of) the denotations can be computed. For object-denoting expressions, this amounts to finding a link (or, ranking all possible links) to an object file. Finally, I also assume that a continuous representation for the expression is computed recursively (for example, through recursive neural networks, (Socher et al., 2014)), out of the continuous representations for the words / concepts.

This setup mirrors the tripartition in the concept files. What are the uses of these representations? I have already mentioned the role of the logical form in guiding the computation of (visually referential) denotations. Continuous representations, on the other hand, can provide a fast route to the computation of inferential relations between expressions (along the lines of the recently popular computational task of "natural language inference", (Bowman et al., 2015)). This could, for example, support the fast computation of coherence relations when integrating new content into the discourse (as is assumed by SDRT (Asher and Lascarides, 2003); and, as question-relevance, by KoS (Ginzburg, 2012)). In slower, explicit justifications of such judgements, the logical form can be used in sound, formal reasoning.[12]

Implementation In some preliminary experiments (to be reported in full elsewhere), we have parsed 50,000 captions from the COCO corpus (Lin et al., 2014), using the English Resource Grammar (ERG) (Flickinger, 2011) executed by the Answer Constraint Engine (ACE) (Packard, 2013) and accessed via pyDelphin.[13] In a task that uses captions from the test set as binary questions ("is this CAPTION?") paired with correct and incorrect images, we found that a reference computation approach based on the words-as-classifiers model can work satisfactorily. For a task that asks a model to predict whether two captions label the same image or not (used as a very rough proxy for entailment pairs), we computed representations for the captions using a TreeLSTM (Tai et al., 2015), and again found this to work satisfactorily.[14]

Extensions Before we move on, we note two possible further extensions of the model. First, it might be useful to allow the creation of object files

[11] We abstract here from the fact that a full model of situated discourse processing would need to work incrementally and hence this structure building would have to happen incrementally. However, we do think that the proposal here is broadly compatible with an incremental processing model such as for example that of Schlangen and Skantze (2009).

[12] It is tempting to see an approximation of Kahneman's (2011) System 1 / System 2 distinction here. How far this carries must be explored in future work.

[13] https://github.com/delph-in/pydelphin

[14] There is not enough space here to properly report these experiments, and so I won't attempt to report numbers here (which would not be interpretable anyway without the further detail). I only offer this information as anecdotal evidence that the approach could potentially be made to work.

also based on the discourse representation and not driven by perception. This could be used to model the state Chris is in after (3, "we saw a man carrying a dog") in our example in the introduction, where she can form a mental representation of the situation that is described. Schlangen (2019) describes a simple retrieval-based mechanism for constructing such imaginations of situations and their use in computing bridging relations (e.g., imagining "leash" when "dog" is mentioned). This, at least prima facie, seems to relate to the idea of understanding as simulation (Barsalou, 2012); to what extend needs to be explored in further work.[15] Interestingly, recent computational work on image synthesis from text also makes this distinction between predicting layouts (our object files) and predicting, based on that, the actual images (e.g., (Hong et al., 2018)).

The second extensions builds this out into the other direction (looking at Figure 1). We note that a composed expression representation (e.g., "the first president of the United States") could also be used to guide the recall of *entity files* from memory, via their stored semantic attributes. This might potentially address the definite description variant of Frege's puzzle (which we discussed above with names only), by making it possible for an agent to know of George Washington and still not retrieve the appropriate file via the description ("first president ..."), hence blocking substitution in a believe attitude report.

6 Back to The Story

We can now briefly return to the story from the introduction, and see how agents modelled according to the above would work here. Processing the various utterances involves the creation and linking of different kinds of mental representations, some of which stand in direct contact to the outside world (via visual indices), some of which are connected to memories of previous experiences. A proper test of this model would be an implementation (perhaps in simulation), showing that similar behaviour can indeed be created.

[15] This might even go some ways towards explaining the vast differences in how some people report phenomenal experience when imagining a scene (which in this model would mean "project back" the object files and their classifications into near-experiential impressions), while others lack this completely, but still seem to be able to simulate spatial relations. See Richardson (1999) for a recent overview of research on mental imagery.

7 Related Work

I have already mentioned in passing that there is a large amount of work that is related to the individual components which are brought together here. As a whole, the model probably has the most similarities with models from robotics on the one hand (e.g., (Kelleher et al., 2005; Kruijff et al., 2012), see (Tellex et al., 2020) for a recent overview), where multilayer (discourse and context) models are a salient choice. The literature is vast here, and the exact connections still need to be explored (see also older work on multimodality, e.g. (Luperfoy, 1992)). While the model likely re-invented parts already present in older work, some of the technical components that we aim to bring together are newer than most of this work, and together with the focus on learning, we think, bring a somewhat fresh perspective. On the formal side, the linking approach seems related to the idea of *anchoring* that has been discussed with respect to discourse representation (e.g., (Zeevat, 1999)).[16]

We also think that our proposal is broadly compatible with (and complementary to) approaches that more closely look into the role of coherence relations in situated discourse (Stone et al., 2013; Hunter et al., 2018), but this needs to be worked out. Lastly, the proposal also seems broadly compatible with the long-term efforts of the Gothenburg school (Larsson, 2015; Cooper, forth) to formally describe situated meaning, as well as with the dialogue model of Ginzburg (2012), but puts the emphasis more on the computational side.

8 Conclusions

I have outlined a model of how situated agents can deal with objects they encounter, in the environment, and in their discourses. It brings together elements from more formally oriented work (discourse representation, update), with methods from more processing oriented work (classifiers, distributed representations), in what hopefully at least roughly resembles a coherent whole. Many of the details are still missing, but a framework for where they can be filled in is provided.

Acknowledgements I have presented an even less fully developed version of this as part of an invited lecture series at the LabEx EFL Paris in 2019 (albeit with more details on

[16] This idea has already been present in the early DRT work (Kamp and Reyle, 1993); it is prominent in current unpublished work by Hans Kamp (2018) (as pointed out to me by one of the reviewers).

experiments), and I thank the audiences there for their patience and feedback.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116(3):463–498.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Lawrence W Barsalou. 2012. Situated Conceptualization: Theory and Application. In Y Coello and Martin H. Fischer, editors, *Foundations of embodied cognition*, pages 1–17. Psychology Press.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, USA.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Susan Carey. 2009. *The Origin of Concepts*. Oxford University Press, Oxford, UK.

Robin Cooper. forth. *From Perception to Communication: An analysis of meaning and action using a theory of types with records*. na.

Donald Davidson. 1986. A coherence theory of truth and knowledge. In Ernest LePore, editor, *Truth and Interpretation, Perspectives on the Philosophy of Donald Davidson*, pages 307–319. Basil Blackwell.

Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ryan Farrell, Ramesh Raskar, and Nikhil Naik. 2018. Improving Fine-Grained Visual Classification using Pairwise Confusion. In *European Conference on Computer Vision (ECCV 2018)*, pages 1–17.

Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In Jennifer E. Arnold and Emily Bender, editors, *Language from a Cognitive Perspective: Grammar, Usage, and Processing*. CSLI Publications, Stanford, CA, USA.

Jerry A. Fodor and Zenon W. Pylyshyn. 2015. *Minds without Meanings*. MIT Press, Cambridge, Massachusetts, USA.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Ztschr. f. Philos. u. philos. Kritik*, NF 100:25–50.

Jonathan Ginzburg. 1995. Resolving questions I. *Linguistics and Philosophy*, 18:459–527.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, UK.

Irene Heim. 1983. File Change Semantics and the Familiarity Theory of Definiteness. In R. Bäuerle, Ch. Schwarze, and Arnim von Stechow, editors, *Meaning, Use and Interpretation of Language*, pages 164–189. De Gruyter, Berlin, Germany.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard de Melo, Claudio Gutiérrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge graphs. *ArXiv*, abs/2003.02320.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In *CVPR 2018*.

Julie Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11(10).

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. 1992. The Reviewing of Object Files: Object-Specific Integration of Information. *Cognitive Psychology*, 24:175–219.

Hans Kamp. 2018. Entity Representations and Articulated Contexts: An Exploration of the Semantics and Pragmatics of Definite Noun Phrases. Retrieved 2020-06-15.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

Frank C. Keil and Jonathan F. Kominsky. 2015. Grounding concepts. In (Margolis and Laurence, 2015), chapter 24, pages 677–692.

J. Kelleher, F. Costello, and J. Van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102.

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. Association for Computational Linguistics.

Geert Jan Kruijff, Miroslav Janíček, and Hendrik Zender. 2012. Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems*, 27(2):27–35.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 36(3):453–465.

Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211—-240.

Staffan Larsson. 2011. The TTR perceptron : Dynamic perceptual meanings and semantic coordination. In *semdial 2011*, pages 140–148.

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

Staffan Larsson and Jenny Myrendal. 2017. Towards Dialogue Acts and Updates for Semantic Coordination. In *Formal Approaches to the Dynamics of Linguistic Interaction*.

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.

Peter Ludlow. 2014. *Living Words*. Oxford University Press, Oxford, UK.

Susann Luperfoy. 1992. The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of ACL 1992*, pages 22–31.

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

Eric Margolis and Stephen Laurence, editors. 2015. *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press, Cambridge, Massachusetts, 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*.

Scott McDonald and Michael Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, pages 3111–3119, Lake Tahoe, Nevada, USA.

Adam Morton. 1980. *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Clarendon Press, Oxford, UK.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA, USA.

Woodley Packard. 2013. The answer constraint engine (ace). Software package.

John Perry. 2012. *Reference and Reflexivity*, 2nd edition. CLSI Press, Stanford, CA, USA.

Hilary Putnam. 1975. The meaning of meaning. In *Mind, Language and Reality; Philosophical Papers Volume 2*. Cambridge University Press, Cambridge, UK.

Zenon Pylyshyn. 1989. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32:65–97.

Franois Recanati. 2012. *Mental Files*. Oxford University Press, Oxford, UK.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

John T. E. Richardson. 1999. *Imagery*. Psychology Press, Hove, UK.

Eleanor Rosch. 1978. Principles of Categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum, Hillsdale, N.J., USA.

Ivan Sag and Thomas Wasow. 1999. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford.

David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

David Schlangen. 2019. Natural language semantics with pictures: Some language & vision datasets and potential uses for computational semantics. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Gothenburg.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 710–718, Athens, Greece.

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.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*.

Elizabeth Spelke. 1994. Initial knowledge: six suggestions. *Cognition*, 50(1-3):431–445.

Elizabeth S. Spelke and Katherine D. Kinzler. 2007. Core Knowledge. *Developmental Science*, 10(1):89–96.

Matthew Stone, Una Stojnic, and Ernest Lepore. 2013. Situated utterances and discourse relations. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55.

Sina Zarrieß and David Schlangen. 2017a. Deriving continous grounded meaning representations from referentially structured multimodal contexts. In *Proceedings of EMNLP 2017 – Short Papers*.

Sina Zarrieß and David Schlangen. 2017b. Is this a Child, a Girl, or a Car? Exploring the Contribution of Distributional Similarity to Learning Referential Word Meanings. In *Short Papers – Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

Henk Zeevat. 1999. Demonstratives in discourse. *Journal of Semantics*, 16(4):279–313.