L's final two steps and look at a more formal approach to resolving references using a semantic formalism: *first-order logic*, which allows us to bridge the gap between a conceptualisation of the model and an implementation of it. We will see that first-order logic has some limitations when applied to a practical model/component, but those limitations are addressed in the section that follows by ideas from *grounded* semantics–thus all three of L's steps are addressed and a component can then be modelled.

Of lesser focus in this thesis, but not necessarily of lesser import, is Section 2.5 which explores relevant philosophical literature on reference and meaning, which will help the models fit into the larger scheme of ideas. We return to some of these ideas throughout the thesis. As is shown in Chapters 5 and 6, the presented models of reference resolution give credence to some of the fundamental theories on reference and meaning presented. Finally, this chapter concludes with a listing of several assumptions we must make in order for the model/component to be realised.

2.1 Incremental, Situated Spoken Dialogue Systems

In this section, we look at the context in which reference takes place: spoken, situated dialogue. In the example at the beginning of this chapter, this includes both S and L's contribution to the (albeit short) dialogue.

2.1.1 Spoken Dialogue Systems

As explained in Chapter 1, a *spoken dialogue system* (SDS) is a computational agent that can converse with human beings through everyday spoken language (see Lison (2013)). The most basic form of a SDS requires some way of representing the human user's utterance, usually by attempting to transcribe the speech into written words through an automatic speech recogniser (ASR). That (attempted) transcription is then fed into a natural language understanding (NLU) component that abstracts over the transcribed utterance (e.g., via a syntactic or a semantic representation). That representation is then given to a dialogue manager (DM) which determines the next action to take (e.g., ask some kind of clarification request, or look up requested information in a database). The action often results in the need to utter something back to the human user, which is the job of the natural language generation (NLG) component (e.g., generation of the response utterance, then the actual synthesizing of that utterance). This procedure is visually represented in Figure 2.2.

This kind of SDS setup is typical of the kind of dialogue that would take place over a phone where the only modality of communication between two participants is speech, which is what is typically under focus in SDS research. There is a connection between the dialogue manager

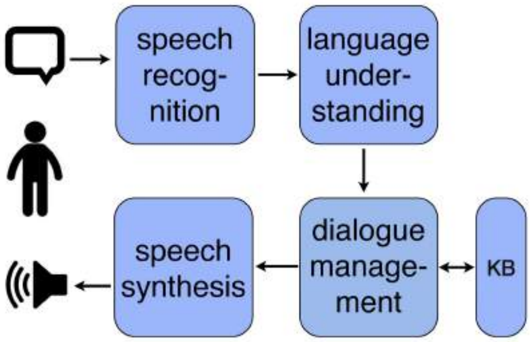

Figure 2.2: Example of a standard spoken dialogue system: when a user speaks, a speech recogniser provides input to a language understanding module, which creates some kind of abstraction over the input that is given to the dialogue manager, which makes a decision on an action to take (potentially based on information from a knowledge base), which generally involves the generation of speech to the user.

and some kind of *knowledge base* (KB) which is a set of facts about the world that the SDS can say something about. For example, if the SDS is a phone-based system that aids users in finding information about booking airline flights, the knowledge base would need to have flight information; e.g., origins and destinations, flight times, information about airports, etc. When a user speaks into his phone, that signal is transmitted and received by the system and passed through the ASR, which gives its hypothesis to the NLU which attempts to form a meaning representation over the utterance. That is given to the DM which determines the best course of action given that utterance, for example, if a request for flight information from Chicago to Atlanta is asked for, the DM would need to look up that information in the KB and then inform the NLG to produce an answer based on the results of that information. These types of systems, and the individual components that make up these kinds of systems, have been well-studied. For example, the *Air Travel Information System* (ATIS) corpus (Hemphill et al., 1990; Dahl et al., 1994) was produced to develop and evaluate NLU. In Chapter 3, we look closer at the NLU component, related literature, and how resolving references relates to NLU.

2.1.2 The Reference Resolution Component

In terms of a SDS component, reference resolution (RR) is the task of resolving REs to the referent. At its highest level of abstraction (that takes all three of L's steps as outlined in the example in the beginning of this chapter), this can be formalised as a function f_{rr} that, given a representation U of the RE and a representation W of the (relevant aspects of the) world (which can include aspects of the discourse context), returns I^*, the identifier of one the objects in the (non-visual) world that is the intended referent of the RE:

$$I^* = f_{rr}(U, W) \tag{2.1}$$

This function f_{rr} can be specified in a variety of ways. Recent work has used stochastic models using the following approach: given W and U, the goal of RR is to obtain a distribution over a specified set of candidate entities in that world, where the probability assigned to each entity represents the strength of belief that it is the referred one. The referred object is then the argmax of that distribution:

$$I^* = \underset{I}{\operatorname{argmax}}\, P(I|U, W) \tag{2.2}$$

A RR component could replace the NLU component depending on the task, or it can be a sub-component of NLU, performing the difficult task of resolving references while NLU handles other processes that produce semantic abstractions over utterances (which could also be useful to a RR component).

As useful as these systems can be practically, as well as in terms of researching how language is used, they aren't sufficient to handle the types of phenomena in dialogue that this thesis explores. As noted in Chapter 1, we need to handle the fact that (1) the space is shared as the participants are co-located, and (2) the time is shared as the participants fluidly take turns and comprehend utterances as they unfold. The standard SDS in Figure 2.2 is not fully amenable to these conditions. In the following sections, we look at the variants of SDS that *are* amenable to these constraints; for situated SDS, and for incremental SDS.

2.1.3 Situated Spoken Dialogue Systems

In a situated SDS, speech isn't the only modality used for communicating between the system and the human participant. As noted in Chapter 1, situated dialogue denotes co-location; the participants can see and hear each other, and they are able to see objects in their shared space. A SDS that can replace one of the participants in such a setting needs to go beyond just processing speech; it also needs to have a representation of the visual situation which, following

Equation 2.1, we call the world W, and it has to be able to observe non-linguistic, yet communicative cues from its interlocutor, namely (for the interests of this thesis) pointing gestures. In other words, the SDS needs some notion of *situational awareness*.

Figure 2.3 shows visually how such a SDS might look: as before, there is ASR and speech synthesis, but the NLU component (which is now a RR component) also has information about the interlocutor's pointing gestures (denoted as *Deixis*), and the KB is replaced by a representation of the world W (though the manner of the representation of the KB and W could be the same).[1]

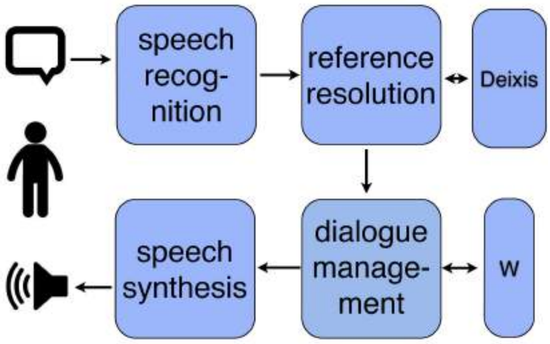

Figure 2.3: Example of a situated dialogue system that has a representation of how the world (W) is represented (similar to the KB in Figure 2.2), also there is a way of recognising the pointing gestures of the human participant.

The representation of W here deserves some additional attention. If goal of RR is to resolve REs made to an object that exists in the immediate space–a visually present object–then a component that resolves RRs needs to know about those objects. That is, W needs to somehow represent individual objects that appear in the scene (how this is done is explained in later chapters, and depends on the way the world is grounded with the language). Those objects are distinct from each other as they are represented, and some kind of visual information about them (either raw features or a computed set of visual properties) also needs to be present.

The component that gives deictic information to the RR component also deserves some additional explanation. Such a module has a fairly big job to do: it must somehow take information about the human interlocutor, and determine not only if that human is performing a pointing gesture, but to what object that human is pointing, or at least some kind of approximation thereof.

[1] One could see a situated dialogue system as one that doesn't just have ears, but also has eyes that can decipher visually present objects and decipher what the human interlocutor is doing.

Modelling W and a deictic gesture are both non-trivial tasks and add to the uncertain nature of the ongoing dialogue. However, they are necessary in a dialogue system that is situated, which is the kind of dialogue that we want the RR component to be able to handle.

Motivation Situated dialogue systems are essential to tasks that require a system to have some awareness about the immediate situation. For example, when driving a car, research shown that when the driver speaks to another person using hands-free devices (e.g., hands-free cell phones) there is a decrease in driving performance (He et al., 2013; Horrey and Wickens, 2006; Ishigami and Klein, 2009; McEvoy et al., 2005). Even just listening to speech causes increased cognitive load on the driver (Demberg et al., 2013). This is not the case, however, when drivers speak to passengers (Drews et al., 2008a), as passengers are aware of the driving situation and can adopt strategies that help the driver perform both the primary task of driving and of the dialogue with the passenger (Drews et al., 2008b). It was shown in Kousidis et al. (2014); Kennington et al. (2014b) that a situationally aware (i.e., situated in that the road and driving conditions were accessible to the system) SDS does not inhibit the driver's driving abilities because the system can respond to situations that require the driver's attention by interrupting its own speech (which also requires incremental processing, which is explained below).

Another practical area is robotics. When embodied robots interact with their environment, they can be made to interact with humans through speech. The type of SDS they would need is a situated SDS, as a robot would need to have some kind of representation of its surroundings, including information about human interlocutors. See, for example Chai et al. (2014); Kennington et al. (2014a).

2.1.4 Incremental Spoken Dialogue Systems

Another aspect of situated dialogue that a standard SDS doesn't typically handle is the fact that dialogue occurs in real-time and is highly interactive. Consider the following example:

(1) a. J: So Sarah ... I hear she has a new dog
b. K: yeah
c. K: She does. In fact, I was there when she bought it
d. J: oh?

J brings up the topic of Sarah's potential new dog. K gives *feedback* (i.e., a verbal indication of understanding; e.g., a back channel) in line (1-b) for J's utterance in (1-a) but does not attempt to take the floor (i.e., take a turn as the speaker). Such feedback is a way for K to signal to J that she has understood his utterance until that point. J also gives feedback to K on line (1-d) for the utterance in line (1-c) for the same reason.

If K is replaced with a SDS, that system would need to produce similar behaviour as K where the feedback is produced as the utterance is ongoing. Traditional SDSs, such as one represented by Figure 2.2, don't work in this way. Traditional systems usually take as input full utterances from the ASR, requiring the ASR or some related component to determine when speaking has begun, and when a silence of specific duration has been detected (this is known as *end-pointing*). When such a silence is detected, the ASR hypothesis is then given to later modules; e.g., NLU. This results in a kind of strict turn-taking style of dialogue between a human and a SDS which can be compared to playing a game of *ping pong*.

Dialogue by its very nature is incremental in that participants in a dialogue take turns speaking (Schlangen and Skantze, 2011). However, in contrast to a traditional SDS, an incremental SDS doesn't wait until the end of an utterance to begin processing, making the increments of dialogue more fine-grained. In principle, an incremental SDS attempts to process as much as possible as early as possible, while attempting to not re-compute parts that have already been computed (more on this below). In reality, an incremental SDS processes the input utterance word-by-word, which has been shown to be a level of granularity in which humans interpret utterances (Brennan, 2000; Schlesewsky and Bornkessel, 2004).

Motivation On a practical level, dialogue systems that process incrementally produce behaviour that is perceived by human users to be more natural than systems that use the traditional turn-based approach (Aist et al., 2006; Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 1991; Asri et al., 2014), offer a more human-like experience for the human users (Edlund et al., 2008) and are more satisfying to interact with than non-incremental systems (Aist et al., 2007). Psycholinguistic research has also shown that humans process (i.e., comprehend) utterances as they unfold and do not wait until the end of an utterance to begin the comprehension process (Tanenhaus and Spivey-Knowlton, 1995; Spivey et al., 2002). This has ramifications for resolving REs: as a RE unfolds, a component that resolves REs should attempt to resolve the referred object at each word increment prefix, updating the belief over candidate referred objects as additional words are added to the prefix.

Work has been done in incremental processing in many areas of dialogue systems: speech recognition (Baumann et al., 2009), speech synthesis (Buschmeier et al., 2012), and dialogue management (Okko et al., 2010; Selfridge and Arizmendi, 2012). Architectures for incremental dialogue systems have been proposed (Schlangen and Skantze, 2009, 2011) and incremental toolkits are also available (Baumann and Schlangen, 2012). More relevant to the work in this thesis is a recent attempt to identify the requirements for incremental semantics in dialogue processing (Hough et al., 2015), as well as work in incrementally processing utterances to produce syntactic as well as semantic abstractions (Demberg and Keller, 2008; Purver et al., 2011; Peldszus et al., 2012; Peldszus and Schlangen, 2012; Beuck and Menzel, 2013). A review of work in incremental RR and related tasks is given in Chapter 3.

2.1.5 Incremental Computation

Comprehending an utterance (or, more specifically, a RE) incrementally is more than just applying a particular component on finer-grained prefixes, such as words. Following Schlangen and Skantze (2009, 2011) we distinguish between two kinds of incremental processing: *restart* incremental and *update* (i.e., fully) incremental. In a restart-incremental system, all internal state is thrown away between updates and output is always (re-) computed from scratch using the current input prefix–not just the newest increment of it. An update-incremental system keeps its internal state between incremental update steps, enriching the internal state at each incremental update with the delta between the current and the previous increment.

The difference between restart- and update-incremental approaches is illustrated in the following two examples (as an incremental ASR component might produce):

(2.3) the

(2.4) the red

(2.5) the red circle

(2.6) the red circle next

(2.7) the red circle next to

(2.8) the red circle next to the

(2.9) the red circle next to the green

(2.10) the red circle next to the green circle

In the above example, as input is received incrementally, a restart-incremental RR component would use the prefix at each increment and recompute what has already been computed. There is no maintenance of internal state. For example, a SDS component described in DeVault et al. (2009) is a NLU component that produces an entire semantic representation (in this case, an expected frame), even if it is only from partial input and no internal state between update steps is kept. Contrast that with the following:

(2.11) the

(2.12) red

(2.13) circle

(2.14) next

(2.15) to

(2.16) the

(2.17) green

(2.18) circle

The above example maintains an internal state and updates that internal state based on new information, without recomputing information that has already been computed.[2]

Of course, processing update-incremental SDS is not a trivial task. For example, the input given by the ASR might not be reliable as it processes incrementally, e.g., it produces output in the middle of a word, and would need to somehow undo the fact that it produced an early hypothesis and then produce the output that is more informed. There are other details that need attention when approaching incremental dialogue that works update-incrementally. In the following section, we look into a recently developed framework of incremental dialogue that addresses these concerns in a systematic model introduced in Schlangen and Skantze (2009, 2011), which plays an important role in deriving the model of RR in Chapter 5, and gives us some concepts and notations that is used in forthcoming chapters.

2.1.6 The IU Approach to Incremental Dialogue Processing

The basis of the model presented in Schlangen and Skantze (2009, 2011) is the *incremental unit* (IU) which is a minimal amount of 'characteristic input' that modules take in, update their internal state based on that input, and in turn produce their own IU output. (This model is often called the IU-model of dialogue processing, and we will henceforth refer to it as such.) This 'characteristic input' can be defined to be anything that is necessary to a particular module; such a definition implies that the granularity is also specified.

[2] The update-incremental approach has obvious benefits such as only needing to update the internal state based on the delta between an increment and the previous increment. However, in Khouzaimi et al. (2014) the authors show that non-incremental components can be made incremental, albeit restart-incremental, which is sometimes preferable over re-modelling and re-implementing a component from scratch to work update-incrementally.

For example, a typical, traditional SDS would define the characteristic input from an ASR module to a RR module to be an entire utterance. Thus, an IU that is outputted from ASR which would then be input to RR would be some kind of representation (e.g., transcription) of an entire utterance. An incremental SDS which typically works on the word level would be finer grained: an ASR module would produce IUs on the word-level; as words are recognised, they are passed to the RR module which would need to be able to process (e.g., update its belief state as to which object is being referred) at each word. This simple, yet important difference is illustrated in Figure 2.4.

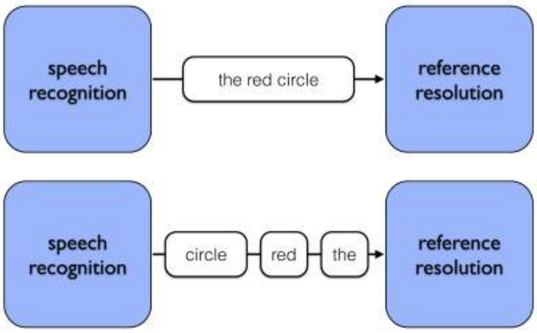

Figure 2.4: Example of the difference between (top) RE- and (bottom) word-level increments being sent from an ASR to a RR module.

How these IUs are defined is up to the system designer. For example, one system's ASR might produce an IU when it determines that a recognisable word has been uttered, whereas another system might produce an IU at specific time increments (i.e., by something other than linguistic units).

Additionally, the words in Figure 2.4 that make up the incremental input are part of a larger utterance and that utterance, in turn, is part of the larger dialogue. This whole-part relation is considered in Schlangen and Skantze (2009, 2011). That is, IUs that are created by the same module can be connected via a relation called *same-level links* (SLLs) which give a particular IU a direct link to its successor and a link to the IU that is its successor. For example, in Figure 2.4, the IU with the word *red* as its characteristic input is the successor of the IU for *the* and the IU for *circle* is the successor of IU for *red*.

Each module in the SDS take a specific type (or multiple types) of IU as input and produce, in turn, its own specific type of IU as output. Thus one module might receive very fine-grained input, but produce output at more spaced intervals. For example, an ASR module might produce IUs at each word, and the NLU module might update its semantic abstraction over the words it has already received at each word. However, the DM, which receives IUs from the NLU, might not produce an action at each individual word. Rather, it might produce a back channel to signal ongoing comprehension (e.g., *m-hm*) after certain words, and produce an action when there is enough information to justify it (e.g., looking up an answer in a knowledge base, or extending a robot's arm to reach for an object that the RR module thinks being described).[3] This gives rise to another important relation: how do the IUs that a module outputs relate to the IUs which that module took in as input? In the IU-model framework, such a relation is the *grounded-in* (GRIN) relation.

A more complete illustration of SLL links and GRIN links is given in Figure 2.5. In this example, there are three modules: ASR, a part-of-speech (POS; i.e., produces a linguistic tag for each word) module which receives ASR output as input, and a module that determines when to produce feedback which receives POS output as input. For each word IU produced by the ASR module, there is an individual corresponding POS IU, but it took three of these POS IUs before the feedback module produced an IU; the feedback IU is GRIN to all of the POS IUs.

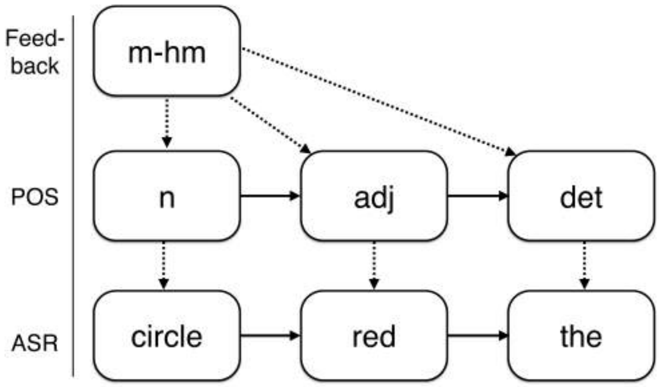

Figure 2.5: Same-level links (solid arrows) provide links between IUs of the same module and grounded-in links (dotted arrows) provide links to the IUs that justify that IU. Time increases to the left.

It is important to note that there are now two networks at play: a network of processing modules, and a network of IUs that were produced by those modules. These will be distinguished as the IU-modules and the IU-network, respectively.

One criticism against using incremental units that are finer grained than utterance level is that, often, waiting for more input means more informed hypotheses. This is certainly the case

[3] Though the decision to produce a back channel certainly is an action.

for ASR, where a word could be recognised, but at some point it is actually found to be a prefix of an ongoing word. This, however, isn't a strike against processing dialogue incrementally as long as there are provisions for handling these cases. The IU-model of Schlangen and Skantze (2009, 2011) defines provisions for such cases where a module can 'change its mind' in light of new information. Specifically, following Baumann and Schlangen (2012), there are three types of operations that a module can perform on an IU (as it pertains to the IU-network), namely `ADD`, `REVOKE`, and `COMMIT`. Each will be explained. Before we do, however, we need to establish how IUs are communicated between IU-modules.

Each module in the IU-modules has three parts: a *left-buffer*, a processor (i.e., an internal state) and a *right-buffer*. An IU-module takes in IUs on its left buffer. When IUs appear on its left buffer, it takes those IUs and processes them, updating its internal state, and produces (when necessary) corresponding output IUs which are put onto its right buffer. Two modules are connected by their buffers: a module's right (i.e., output) buffer can be connected to another's left (i.e., input) buffer. Thus when a module produces new IUs, those IUs are simultaneously put onto its right buffer, and any other module's left buffer that is connected to that module's right buffer. This set of IU-modules and the connections between them form the IU-module network, whereas the IU-network is a network of the IUs themselves. We now return to the explanation of the operations on IUs that modules can perform.

ADD Adding an IU is the operation that adds an IU onto the IU-network. When a module takes in new input on its left buffer and processes that input, it may produce its own characteristic output which is packaged into an IU. A module has the obligation to do several things when an IU is produced. First, the SLL and GRIN links must be established, giving the IU place in the IU-network. Second, the modules that are dependent on that module's output need to be informed that the IU has been added to the IU-network; i.e., the modules whose left buffers are linked to that module's right buffer need to be informed that an IU has been added. Those modules can in turn act upon that new input.

REVOKE As a module receives new input, it may determine that an IU that it had previously produced is no longer valid, and as a result should no longer be a part of the IU network (and, usually, would be replaced with another IU). When this occurs, the module has an obligation to inform the modules connected to its right buffer that a particular IU has been revoked. It could be the case that the IU has already been processed in later modules, so a module that receives a notification that an IU has been revoked needs to update its own additions to the IU-network that were based on that IU, which would result in the obligation for that module to inform the modules that are connected to its right buffer that an IU has been revoked, an so on. In general, once something has been revoked, the new IU that “takes its place” can be added to the network as described above.

COMMIT When a module determines that an IU has been added to the network will not, by any circumstance, be removed from the network, it can inform modules that are connected to its right buffer of this decision. For example, when an ASR module determines that an interlocutor has stopped speaking, there will be no additional input, so the hypothesised transcription in the form of IUs that it has produced will not be revoked (i.e., there will be no new information to inform such a revoke). No additional operations are needed to augment the IU-network, but the IU that has been committed now has the state of being committed, and the other modules must be informed of this change in the IU-network. This might be useful information to later modules that have to make a decision; it would be informative to know that a decision can be made based on what is already in the IU-network–waiting for more information would not be beneficial.

Example The ADD and REVOKE operations are illustrated in Figure 2.6 for the RE *the red circle*. As the ASR module adds new words to the IU-network, the POS module is informed of each added IU which gives rise to that module's corresponding input POS IUs. During processing, the ASR adds the word IU for *sir*, but determines later that *sir* was actually the beginning of the word *circle*. The IU for *sir* is revoked, and the ASR module informs the POS module that a revocation of the IU for *sir* took place. The POS module then revokes its own IU for *n* (noun) which was produced (i.e., GRIN) by the word IU for *sir*. The IU for the word *circle* is then added to the network, which informs the POS module, which produces an IU that grounds into the IU for the word *circle*. The rest of the process continues with ADDs (and COMMITs if the ASR detects a certain amount of silence).

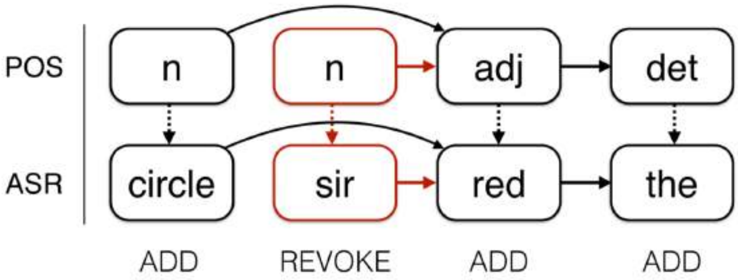

Figure 2.6: Two modules perform ADD operations until *sir* is replaced by *circle* using a REVOKE, then an ADD operation. The ASR module informs the POS module of the changes, and the POS module updates its corresponding IUs accordingly.

Summary of the IU-Model Following the explanation given in Kennington et al. (2014d), SDSs based on the IU-network approach consist of a network of processing *modules*. A typical module takes input from its *left buffer*, performs some kind of processing on that data, and places the processed result onto its *right buffer*. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules. The IUs themselves are also interconnected via *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect it. A complication particular to incremental processing is that modules can "change their mind" about what the best hypothesis is, in light of later information, thus IUs can be *added*, *revoked*, or *committed* to a network of IUs.

More information on the IU-model of incremental dialogue processing can be found in Schlangen and Skantze (2009, 2011). The authors also note that the traditional SDS is a special case of the incremental SDS where the granularity is simply at the utterance-level instead of the word level. The IU-model has also been adapted into a practical toolkit presented in Baumann and Schlangen (2012) (which was extended to make situated dialogue more streamlined in Kennington et al. (2014d)), allowing practical systems to be built using the IU-model as a framework ranging from in-car dialogue systems (Kousidis et al., 2014; Kennington et al., 2014b) to a dialogue system that a robot would use to play a game with human participants (Kennington et al., 2014a).

The models presented in Chapters 5 and 6 and the components that they are implemented as (see Appendix A and B) make use of this incremental processing model. They work in an update-incremental fashion, which requires them to be able to handle `ADD`, `REVOKE`, and `COMMIT` operations.

This section provided an explanation of SDSs and how a component of RR fits into that, with emphasis on situated, incremental variants of SDS. With this, we now isolate what the RR component is doing (i.e., what L did in the example at the beginning of this chapter) when it incrementally resolves REs.

2.2 Types of Referring Expressions

In the previous chapter, we showed that REs are a type of NP. In this section, we look at the types of NPs that make up the REs that we focus on in this thesis, and, to make matters clearer, some types that are not considered. We begin with the latter.

2.2.1 Some Types we don't Consider

To illustrate, consider the dialogue in (2)

(2) a. J: Did you hear that Sarah has a dog?
b. K: Yes, I was there when she bought it.
c. J: Ah, so the dog is real.
d. K: Yes, yes, her dog's name is Biff.

Proper Names

Biff in (2-d) is an example of a *proper name* usage for the dog under discussion between J and K. In this thesis we do not focus on proper names. The reader is referred to Chapter 5 of Abbott (2010) and the references there for further investigation into proper names.[4]

Sentential Propositions

A proposition is a statement or assertion that expresses a judgement or opinion, as in Example (3).

(3) this sentence is an example of a proposition

In the example dialogue in (2), the statement *the dog is real* made by J in (2-c) is also an example of a proposition, as it appears during the course of a "typical" dialogue. Generally, propositions require some kind of existence verb, i.e., a statement of *identity* (e.g., A *is* B), which equates the two parts of the proposition. In the example in (2-c), J equates *the dog* with being *real*.[5]

[4] This is not to say that naming a visually present object and using that name to later refer to that object is not a phenomenon that happens in a reference task such as we are interested in. This has been looked into by others (Chai et al., 2014) and focuses on establishing common ground between dialogue participants, as in Clark (1996), which is related to learning the kind of meaning we attempt here.

[5] The pattern of a proposition is A is B, something that Burge (2010) called an *indication*, which differs from reference in an important way. Of indication, he said (p.32):

> ...functioning to refer does not constitutively carry with it a function to engage in attribution or functional application. Since attribution is a constitutive representational function of the predicate 'is red' and the concept is red, they do not refer to anything. They indicate the property of being red. A primary representational function of predicates in language and predicative concepts in thought is attribution. So predicates and concepts indicate entities—bear relations to aspects of a subject matter. Their doing so is fundamentally in the service of attribution, attributing such aspects to further entities (often entities that are referred to). In occurrences in which no logical operations, such as negation, are involved, the predicate and the concept function to attribute what they indicate. For example, in That apple is red, is red functions to attribute what it indicates—the property redness, or the property of being red—to what That apple refers to. In attributing a property, they represent something as having that property.

In this thesis, the focus is on definite descriptions that don't necessarily have propositional value. Consider the difference between (4-a) and (4-b):

(4) a. the red circle
    b. the circle is red

The example in (4-a) *presupposes* (that is, assumes) that a visually present object has the properties that are being uttered, namely that it is red and has a circular shape. It is assumed that the notion of 'red-ness' and 'circular-ness' have already been learned and will be understood by the listener. This is in contrast to (4-b) where there is a proposition that an object is red, as if the notion of 'red-ness' is either not yet learned, or there is some kind of question as to the circle's actual redness that needs addressing via a proposition. For more work about propositions, the reader is referred to the papers mentioned in Section 2.5 below.[6] Included in what we will not consider are propositional attitudes where propositions can be prefaced by things like *I think* or *I believe*, e.g., *I think that I want you to look at the red circle* or *John believes that there is a red circle*. The NPs that we are interested in can occur from within those kinds of sentences, but we aren't interested necessarily on those kinds of sentences as a whole.

Indexicals

The resolution of a referring NP that depends on immediate surroundings is, by definition, indexical (Barwise and Perry, 1981, p. 33). In other words, all of the types of NPs that we are interested in are indexical. However, there are some (more prominent) indexicals that we do not consider, which we strictly refer to here as indexical. Specifically, personal pronouns such as *I*, *we*, *you*, *he*, *she* that generally refer to people and the referents of those indexicals are highly dependent on the dialogue context (e.g., who is speaking, who is listening, etc.). Though there will always be a speaker and a listener, and they could be considered as "objects" which are visually present and hence referable, they nevertheless will not be of particular focus in this thesis. For discussion on these kinds of indexicals, see Kaplan (1989); Nunberg (1993). The reader would also benefit from Chapter 10 of Daniels (1990).

2.2.2 The Two Types we do Consider

As mentioned in Chapter 1, in this thesis we are primarily concerned with NPs that refer to an entity that is visually present. Such NPs that require contextual information for resolution are

[6] The RE *the red circle* is arguably a proposition, albeit a very weak one. For example, it is the same to say *the red circle exists*, which is a kind of proposition, but this is also an attributive use (as in Donnellan (1966)) rather than a referential one.

forms of **deixis** (also termed *indexical*, as noted above). The goal of their being uttered is to direct a listener's attention to the referred object.

Examples of these kinds of NPs are in (5). Each type will be introduced in turn.

(5) a. J: What about **the red one**?
b. J: (pointing) **That** one.
c. S: I think **it** is too small.[7]

Definite Descriptions

To put (5) into context, J is referring to an actual, single (more on this singleness constraint in the next section), visually present entity (i.e., an object). He initially uses the NP *the red one*, which is a type of definite description that contains words that pick out visual properties that the object has. Descriptions such as this don't name an object directly, rather aspects of an object. In order to resolve such a reference, the entire RE must be resolved and the lexical meaning of the words that make up definite descriptions must be somehow represented and composed.

Demonstratives

The utterance in (5-b) is an example of a *demonstrative*, deictic word, where some kind of non-linguistic cue is used to refer to the object (e.g., a pointing gesture, or by directing gaze to the object) and the utterance of the demonstrative word (such as *that* or *there*) indicates to the listener that a non-linguistic cue is also presented to aid in the resolution of the reference.[8]

[7] Because it is not a first mention, we are to a lesser degree interested in pronouns, e.g., the pronoun *it*, but we are nonetheless interested in that such a pronoun usage refers *exophorically* to the object. Though it does refer *anaphorically* (i.e., a reference to a previous linguistic entity) to the NP in (5-a), we are more concerned that it refers to a visually present object rather than a linguistic antecedent. In most cases, the kinds of pronouns we are interested in will indeed have a linguistic antecedent, like a definite description as explained above where, having been described sufficiently to direct the listener's attention to an object, subsequent discussion about that object indexes it by a pronoun. However, in all cases the object that is being referenced by a pronoun does physically exist and is visually present, so we will treat it as an exophoric pronoun. Pronouns like this do not draw a listener's initial attention to an object indicated by a speaker as definite descriptions and demonstratives do, and therefore function in a somewhat different way. However, a component of RR should be able to handle them under certain circumstances.

[8] These types of REs also fit into the *Giveness Hierarchy* set forth in Gundel et al. (1993), though how they are used in these differing ways will not be explored. Suffice it here to say that these different types indicate different cognitive statuses (or belief of status) of the speaker, namely, that definite descriptions are uniquely referential, demonstratives are familiar (e.g., both participants can see the object), and pronouns are already "in focus". We can assume that we are focusing on the most restrictive types (4-6), that the other types are subsumed. They found that,

2.3 Modelling Reference with First Order Logic

Section 2.1 set up the context where RR takes place and the previous section isolated the types of REs that will be our focus. In this section, we move closer to actually modelling RR by appealing to a more formal system that has been used in semantics: first-order logic (FOL). We explain how the formalism works, give examples of its application to the types of REs we are considering, and show some of the shortcomings in the kind of model/component that we wish to construct.

2.3.1 Syntactic Assumption

Though we are primarily concerned with the NPs that make up three specific types of REs, they often occur within larger utterances, e.g. (the target REs are shown in bold typeface):

(6) a. I wanted to ask if you can see **that red circle** and tell me what you think of **it**.

Here, and in subsequent chapters, we assume that a component exists that removes NPs that refer from their greater context (e.g., this could be done syntactically by finding NPs and then determining if they refer; see, for example, Friedrich and Pinkal (2015)). Given these NPs that refer to visually present objects, we can now take a closer look at how they could be resolved with FOL.[9]

2.3.2 Definite Descriptions

FOL consists of *terms*. There are several types of terms; the ones we use are *variables*, *predicate* symbols, and *formulas*. We begin with formalising definite descriptions. To illustrate, to say *the red circle* in a more formal way is to say that a circle exists and that circle is red. In other words, there is an entity that exists (more formally, a variable) and it is in the class of circles (a predicate) and it in the class of red things (another predicate). Let's focus first on the circle class:

(7) $\iota x.circle(x)$

The FOL representation in (7) is read as follows: that unique entity x such that x is a circle. In other words, that entity x is in the class of circles. This nicely frames the problem of

the more restrictive the type, the shorter the linguistic content (hence pronouns are generally most strict and have less content). This is not of particular focus of this thesis, but useful for drawing potential insight.

[9] Note that we also assume that the scope of the NP makes up a complete RE. We further assume that we are referring to single entities. More general quantification in the models presented in Chapters 5 and 6 is left for future work, but we provide some thoughts on the matter in the final chapter.

resolving references to resolving the object which can replace x which satisfies the formula; i.e., in which the formula evaluates to true. The variable x can range over all of the objects in a domain (defined shortly) and the ones that satisfy the formula evaluate to true. Clearly this assumes that such a mentioned object exists, and that assumption (indeed, constraint) is applied to the model here; in other words *existence* is presupposed. The ι (iota; read *unique x such that*) symbol expresses the contribution of the definite article which restricts the NP to denote a set with a single entity (Peano, 1897). This is a hard constraint that we are putting on the model; it resolves to one and only one entity, known as the constraint of *uniqueness*. Now that we have a formal representation (i.e., abstraction) over the utterance, it needs to somehow be "connected" to the world. To this end, consider the example in (8):

(8) a. 🟩 $_a$ 🔴$_b$
b. $\iota x.circle(x)$

If the context contains only the two objects in (8-a) whose suffixes represent their object identifiers, we can then range over the object identifiers of the context in (8-a). That is, following model theory (Tarski, 1956; Chruch, 1940), we have a model $M = \langle D, F\rangle$ where $D = \{a, b\}$ (the two object identifiers), and there is a characteristic function which maps the objects in D to the set of circles, $F(circle) = \{b\}$ and another which maps to the set of squares $F(square) = \{a\}$. What we want is to know the unique object identifier $d \in D$ such that the denotation (represented as $[\![]\!]$) of the formula $[\![circle(x)]\!]^{M,g[d/x]} = 1$, where g is a special function that assigns x to d for each possible value of d in the model M. We see that when x is replaced by b, the formula evaluates to true, as in (9) where the uniqueness assumption imposes the constraint that *no more than one* of the formulas evaluate to true; the existence assumption imposes the constraint that *at least one* of the formulas evaluate to true.

(9) a. $[\![circle(x)]\!]^{M,g[d/a]} = 0$
b. $[\![circle(x)]\!]^{M,g[d/b]} = 1$

This works fairly well for single-predicate NPs such as *the circle*. The slightly more complicated FOL formula for *the red circle* given in (10) makes use of the logical *and* operator $\wedge$:

(10) $\iota x(circle(x) \wedge red(x))$

(10) evaluates to true if there is a unique entity in both $circle$ and red that makes the formula evaluate to true, i.e.: $[\![\iota x(circle(x) \wedge red(x))]\!]^{M,g[d/x]} = 1$ iff $[\![circle(x)]\!]^{M,g[d/x]} = 1$ and $[\![red(x)]\!]^{M,g[d/x]} = 1$ and there is only one value that simultaneously satisfies both $circle$ and red. With this, we can derive a simple way of composing the words (which, for the most part,

map directly to predicates) of definite descriptions via the $\wedge$ operator in definition (11):[10]

(11) $[\![a_1 \wedge a_2 \wedge \cdots \wedge a_n]\!]^{M,g[d/x]} = 1$ iff $[\![a_1]\!]^{M,g[d/x]} = 1$ and $[\![a_2]\!]^{M,g[d/x]} = 1$ and ... and $[\![a_n]\!]^{M,g[d/x]} = 1$

This is an *intersective* approach to composition, which is explained in further detail below. We showed that this works for definite descriptions with just the head noun (e.g., *the circle*) and when an additional adjective modifies the noun (e.g., *the red circle*). But the definition in (11) allows us to construct formulas of arbitrary length. Some examples with corresponding FOL representations:

(12) a. *the small red circle*
b. $\iota x(small(x) \wedge red(x) \wedge circle(x))$
c. *the red circle on the left*
d. $\iota x(red(x) \wedge circle(x) \wedge on\_left(x))$

This even covers cases where relational prepositions are used to denote a relation between objects, e.g.:

(13) a. *the circle above the square*
b. $\iota x[circle(x) \wedge \iota y(square(y) \wedge above(x, y))]$

Where that the $above$ predicate takes two arguments, but still maps those arguments into an individual pair (e.g., $F(above) = \{(a, b), (d, c)\}$ where a is above b, and d is above c).

2.3.3 Demonstratives (and Pronouns)

The framework so far works well enough for the kinds definite descriptions are considered throughout this thesis. We now look at formalising pronouns and demonstratives, which aren't completely straightforward. For example (here we are reminded that both of these types of NP assume existence and uniqueness as described above):

(14) a. *that*
b. $\iota x.that(x)$
c. *it*
d. $\iota x.it(x)$

[10] This assumes an identity relation between the application of x that the same value for x is applied to all of the individual functions.

The FOL representations for *that* and *it* in (14) seem intuitive, but the characteristic functions aren't as easy to explain. For it (or other like-pronouns), we can define the characteristic function to map from entities in the domain D to the set of entities that were most recently referred.[11] For *that*, the characteristic function maps from D to the set of entities that are currently being pointed at. Perhaps somewhat more intuitive to these meanings are the following FOL representations, (15-a) for *it* and (15-b) for *that*:

(15) a. $\iota x.recently\_referred(x)$
b. $\iota x.pointed\_at(x)$

2.3.4 Functional Application of Objects

We use an additional operator, the λ-operator, to abstract over variables which are useful when we wish to apply the variables (in this case, objects) in a domain D to a FOL representation (applying objects to formulae in this way is a more direct method of applying a world representation of a formula than using model theory).

$$[\![w]\!]_{obj} = \lambda x.\phi_w(x) \tag{2.19}$$

Where a symbolic concept ϕ_w of a word w (e.g., red or $circle$) have corresponding predicate concepts in FOL and a representation of an object x in D can be applied to that concept via the $lambda$ operator.

With these representations, we can semantically represent the kinds of NPs we are interested in in a simple FOL form, assuming that the domain D and the functions F are fully specified.[12] That is, we assume, if the concepts ϕ_w can be defined and their functional application learned, then the task of fitting application of objects into FOL is more or less complete. The task for this thesis, then, is to couple objects x with language concepts ϕ_w. This is a form of *lexical* semantics, where the "meanings" are considered. More on this below.

[11] This is of course a gross simplification as to what pronouns actually do, including their pragmatic constraints, but we aren't quite concerned with that here; a pronoun will always refer to a visually present object that is somehow salient. We will see examples later when this simple characteristic function doesn't work.

[12] Functional lambda operations were a big part of Montague's (explanation below) semantic (and syntactic) framework (Hobbs, 1983; Cotelli et al., 2007). Here, the only thing applied functionally with lambda are the objects to single predicates. Other operations, such as those connecting those single predicates with other predicates, are defined above in FOL.

2.3.5 Limitations of Intersection

The FOL approach explained above assumes an *intersective* mode of interpretation. That is, as we have defined the *and* $\wedge$ operator, the method of composing the results of the application of two different characteristic functions (e.g., *red* and *circle*) is to find the unique object that exists in both sets (i.e., the intersection; e.g., the object that is both in the set–the class–of red things and in the set of circles). This has its limitations, because some sets are determined relative to each other. Consider the following examples:

(16) a. the small elephant
b. $\iota x.small(x) \wedge elephant(x)$
c. the big mouse
d. $\iota x.big(x) \wedge mouse(x)$

Clearly, these are not composed in the same way as described above, because even a small elephant is larger than a big mouse; i.e., the class of small things doesn't necessarily include the elephant as described; rather, an elephant can be considered small by comparing it to other elephants–the same is true for big mice. This is a limitation of this approach, but we show in Chapters 5 and 6 (and particularly using the data presented in Chapter 4), the assumption that classes are composed in this intersective way works well in practice. We leave more involved methods of composition that handle these kinds of phenomena for future work.

Some Key Shortcomings of First-Order Logic

Even though we are interested in fairly limited phenomena that can be captured with FOL, the work of resolving REs is not yet done. Theoretically, we have a well-established system that we can use to compose a RE in order to determine which object has been referred. However, there are some issues with this approach, particularly when applied to a practical component.

First, is the practical representation of the domain D and the set of functions F in the model M. In order for FOL to work, both must be fully specified. How this is to be done in a practical component constitutes one of the main parts of the task. For D, each object needs to be identified and represented.[13] Perhaps a greater difficulty is in F, that is, defining the characteristic functions that map the objects into specific classes (e.g., red, $circle$, or on_left). In traditional work in RR, there is a simple text-match mapping between a word, e.g., *red* and a

[13] In a system that uses a virtual scene, this is usually quite straight-forward, as objects are already symbolically represented and can be given object identifiers. For a scene that has real-world objects, some kind of vision processing needs to be done in order to segment and represent the individual objects. We look at various approaches, including the approaches presented in this thesis in later chapters.

corresponding predicate red, and the objects in $F(red)$ are defined beforehand. In short, there needs to 1) be a way of determining the set of F, which we also term the *classes* that objects can belong to (e.g., the class of *red* things), and 2) how to determine whether or not an object fits into those classes, i.e., the actual application of each function in the set of F.

Another shortcoming is the assumption (in fact, the requirement) that there is a single referent, *without uncertainty* (i.e., existence and uniqueness). While these are assumptions that we rigidly make here, this poses a problem in practical components which need to be able to handle uncertainty in both the recognition of the RE, and the representation of the world that makes up D and F. For example, someone might describe something as vaguely red, when it doesn't fit the prototypical red that would put that object into the class of red things. Something that isn't prototypically red, but still redder than the other visually present objects should be more likely to be the referred object.

A third shortcoming is the way the REs are composed in a FOL framework. Typically, an entire RE must be specified in FOL before it can be computed and must be "unpacked" from the bottom up, but we noted earlier that humans process the resolution of REs incrementally, i.e., they don't wait until the end of an utterance before processing, rather they process as much as possible as early as possible (which is how we want to model the resolution of REs in this thesis). Example (13) above is a good example of this: standard FOL would need to completely resolve y before it can resolve x. However, in an incremental system each word in the RE should contribute to the resolution to the referred object as the RE unfolds without needing to wait for the resolution of one of the variables (i.e., both are resolved simultaneously, given the unfolding RE).

To recap, the shortcomings that need to be addressed for a practical system are enumerated as follows:

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

These shortcomings constituted the limitations of FOL, but they also constitute the shortcomings of previous approaches to RR (discussed in greater detail in the next chapter). Overcoming these shortcomings is one of the main contributions of this thesis. How two these shortcomings might be overcome is explained throughout subsequent sections of this chapter.

2.3.6 A Brief Survey of Other Semantic Approaches

FOL has been used in logic and semantic theories for a respectable amount of time, and it is clear that FOL doesn't capture all of the phenomena in language use. However, for the purposes of this thesis it is sufficient. Other approaches to representing language semantically exist; indeed ones that are designed for use in dialogue, so why not use those instead of FOL? In this section, we address this important question by briefly describing other approaches to semantics and provide reasoning for using FOL.

Montague Semantics

Richard Montague developed a system of symbolic logic that was introduced in Hobbs (1983); Cotelli et al. (2007), promoted in Partee (1975). Among other things, Montague attempted to solve intensional constructions (see below) and quantificational NPs, however, we don't really have need of generalised quantifiers when dealing with NPs that refer to a single object. That, of course, is *not* to say that they aren't important. We leave other types of quantification to future work. Indeed, many have followed Montague and looked into generalised quantification, see, for example Barwise and Cooper (2008); Hintikka (1973). He also formalised a way to handle the syntax of natural language using categorical types which we could use, but instead we will assume some simple syntactic structure that plays directly into the FOL representations and work from there.

Discourse Representation Theory

(DRT) is a theoretical framework for discourse phenomena such as anaphora and tense (Kamp, 1993). It goes beyond the sentence level, parting ways with formal semantics (e.g., FOL) but does continue to use model-theoretic tools to represent a discourse. Such a framework could have use here, as we are interested in a dialogue setting. However, we aren't interested in the interactive, multi-sentence, discourse aspects of dialogue per se; we are interested in individual REs, the words that make up those REs, how they refer and what they mean. Certainly, DRT can handle phenomena which are difficult to represent in FOL, such as anaphora (in particular, donkey sentences), but they are not needed in this thesis.

Situation Semantics

Situation semantics is a framework that represents the situation of a given speech event. Situations consist of 'individuals having properties and standing in relations at various spatio-temporal locations' (Barwise and Perry, 1981). Those situations can be real or abstract; the former are real situations (like the ones we are interested in) whereas the latter are akin to set-theoretic objects that are constructed from individuals. Closer to what we are interested in, Kratzer (2011), focuses on situations, which has been extended in work by Paul Elbourne (2001, 2008); Daniels (1990).

Of particular interest are his lexical entries for *the*, *it* (Daniels, 1990), and *that* (Elbourne, 2008).[14]

(17) a. $[\![the]\!] = \lambda f.\lambda s : s \in D_s \,\&\, x.f(x)(s) = 1.\iota x f(x)(s) = 1$

b. $[\![it]\!] \quad = \lambda f.\lambda s : s \in D_s \,\&\, x.f(x)(s) = 1.\iota x f(x)(s) = 1$

c. $[\![that]\!] = \lambda x.\lambda f.\lambda g.\lambda s.\iota z(f(x)(\lambda s'.z)(s) = 1 \,\&\, g(\lambda s'.z)(s) = 1$

Note that the entries for *the* and *it* are identical. In other words, in terms of situation semantics, there is no difference between a definite phrase and a pronoun. That is, there is a situation s in a domain D_s, where x is an entity that exists and there is only one such entity (hence ι). These take a NP (minus the determiner) as an argument (e.g., *red circle*). The entry for *that* is trickier, but basically shows the same thing: that there is an entity that exists, that there is only one such entity, and that entity has a gesture (similar to our *pointed_at* predicate in Example (15-b)). The entry for *that* presented in this framework does account for the types of usage in which we are interested, such as using *that* with a pointing gesture, or *that* followed by a definite description (with optional pointing gesture). Models of reference resolution using situation semantics have also been proposed (Poesio, 2011). Situation semantics have been shown to handle phenomena in language that FOL and traditional Montague Semantics cannot on its own (e.g., donkey sentences, see Kratzer (2011)).

Type Theory with Records

Type theory with records (TTR; (Cooper, 2005, 2012)) in a way takes up the intuitions of situation semantics with a different (and arguably less problematic) formalism. TTR is an integration of Montague semantics, DRT, as well as frame semantics (Fillmore, 2006). TTR is well-suited for dialogue. It was the principle framework used in Ginzburg (2012), which was concerned with the issue of how to describe certain linguistic features of interactive conversation. TTR works well because it represents aspects of semantics as we've discussed them, but also utterance types which are aspects of language that are more pragmatic, such as speech acts (Searle, 1976); e.g., clarification requests in dialogue, and other dialogue moves. While these are all important aspects of a fully-functional dialogue system, and we are certainly interested in resolving references within a dialogue framework, the phenomena of dialogue are beyond the

[14] Some aspects of these entries are left out for simplicity.

scope of this thesis.[15]

Discussion

While these other formalisms and frameworks yield richer semantic representations (as well as, in some cases, pragmatic), all build upon FOL in various ways. It is therefore assumed that, if the models presented in Chapters 5 and 6 fit into a FOL framework, then they can be used in extended frameworks such as DRT, situation semantics, and TTR. Importantly, none of these approaches directly overcome the shortcomings of FOL listed above. To begin overcoming those limitations, we now turn to grounded semantics.

2.4 Reference and Grounding

In this section we take a look at the background on *grounded semantics* which provides a foundation upon which we can build as we overcome the some of the above-mentioned shortcomings of FOL. We begin by asking a question that has been asked before (e.g., Roy and Reiter (2005); Larsson (2015)), namely how does language relate to the non-linguistic world and how does linguistic meaning relate to *perception*? How do we as humans learn words and agree on their meaning such that we can use those words to convey and understand our intentions with each other? The formal approaches in the previous section aren't quite able to answer these questions directly (Steels and Kaplan, 1999), and we will see that the answer to these questions help address the shortcomings of FOL outlined above.[16]

In this section, we first look at some approaches to meaning from the field of cognitive science. We then look at how the ideas reviewed in this section fit into reference, particularly in definite descriptions, and demonstratives. We then look at how the ideas in this section can

[15] It should be noted here that attempts have been made at using TTR as a formal basis for lexical semantics (Larsson, 2015). We discuss this further below.

[16] As an aside, the claim that manipulation of symbols, be they computational or logical symbols as presented in the previous section, does not equate to intelligence (in which meaning plays an important role) has a long and interesting debate. One well known contribution was made by John Searle (1980) in his famous "Chinese Room" *Gedankenexperiment*. Searle attempted to make the case that a program cannot give a computer a mind, understanding, or consciousness in the same way that humans do regardless of its behaviour, even if that behaviour seems to be intelligent. In artificial intelligence (AI) research, there is a view of *strong* AI (in which a correctly functioning computer program that behaves human-like is, in fact, intelligent and has a mind in the same way humans do) and *weak* AI (where computer programs simply simulate human behaviour, but don't have human-like minds). In this thesis, an attempt is being made to address some issues that are directly related to AI such as capturing the meaning of words, but we are not making any claim that the models presented here are in fact models of how things are done in a human's mind, nor are they by themselves intelligent. The goal, rather, is to build a component that would be used in a system that a human would interact with in a natural way.

be reconciled with semantics, and extend our formal approach to work with these ideas. We then revisit some discussion from the previous section in light of the ideas explored here.

2.4.1 The Symbol Grounding Problem

Roy and Reiter (2005) make the following claim:

> There is sometimes a tendency in the academic world to study language in isolation, as a formal system with rules for well-constructed sentences; or to focus on how language relates to formal notations such as symbolic logic. But language did not evolve as an isolated system or as a way of communicating symbolic logic; it presumably evolved as a mechanism for exchanging information about the world, ultimately providing the medium for cultural transmission across generations.

Indeed, in the previous section full attention was given to how REs can be represented in a (albeit simplified) formal logic. However, language isn't used in isolation–certainly language isn't isolated when being used to refer to visually present objects. When dealing with language and meaning, we must look at how everyday language is used by humans. The human brain is physically embodied and can interact with its environment through percepts such as sight, sound, touch, etc. Humans also do not exist alone, but are surrounded by objects (as noted in the Chapter 1) with which we can interact, and, importantly, humans also come into contact with other humans. If human A wishes to draw human B's attention to an object that is external to both of them (for example, a piece of fruit, or as in the example at the beginning of this chapter), how can such an intention be performed? Regardless of how language came to be what it is now, the fact of the matter is that humans do use spoken language to communicate with each other in such a circumstance as drawing attention to a piece of fruit. Moreover, the choice of words is important: the individual words in the RE produced by human A must "pick out" properties which human A perceives that particular piece of fruit to have (e.g., its colour, shape, or size) knowing that human B would understand those particular words to pick out the properties of the intended referent. That is, both A and B have a representation of words that pick out visual properties and those words are agreed upon by both. Without such an agreement, communication via spoken language could not occur at all.

With this, we can define what is meant by **grounding**: the agreed-upon meaning of a word is based on perceptual experiences with which that word associates. That is, the meaning of (many) words is *grounded* in perceptual experience. For example, one cannot really know the meaning of the word *red* if one has not seen a red object and experienced that colour word being associated with a visual stimulation in the form of redness.

This seems to be the form of meaning we are after when dealing with reference to visually present objects, but can such a grounded meaning be somehow represented in a symbolic system such as FOL? Harnad (1990) notes that indeed symbolic approaches, such as our formal approach in the previous section, are autonomous functional modules that simply need to be "hooked up" to the world to work. Beginning with a symbolic system such as we have discussed above and hooking it up to the world amounts to representing the world in some way, i.e., mapping from events and entities in the world to a representation that can then be somehow useful to the symbolic system. In our above example, repeated here

(18) $\iota x(circle(x) \wedge red(x))$

resolving the RE which gave rise to the FOL in (18) amounts to finding the entity in the domain D that makes the statement true. But if we are referring to visually present objects, how can we represent those objects such that they can be in D (as explained above)?

This is the problem that we eluded to above: How is it decided that an object fits into the class of red things, or that an object has the property of being red? A rule (i.e., a pre-defined function) could be made where, for example, if an object is perceived as having a value in the RGB scale that falls in the range of things that are generally called red, then that object has the property of being red. This seemingly top-down approach to linking symbols to percepts by defining functions has several problems. First, there are potentially exceptional cases where a non-red object is described as being red, albeit not in the prototypical range of red as defined by the function. This is a problem in robustness. The second problem is that the meaning of a word is encoded in the function–the function is not learned by experience, which is how humans seem to learn meanings of words.[17]

Grounded semantics addresses this problem by learning the functions "by experience" from data. For our purposes, given *training* examples of referred objects and the REs that were used to refer to those objects, a model of RR would need to learn a mapping between object

[17] These problems have been looked into by *connectionist* approaches to cognition (again, for our purposes, we are only interested in cognition as it pertains to word meanings), where neural networks have been introduced as mimicking, to some degree, how the brain works. (See earlier work in connnectionist models (Hinton et al., 1986) and comparisons between symbolic and neural learning approaches (Shavlik et al., 1991). See also Pinker and Prince (1988); Fodor and Pylyshyn (1988) for criticisms to early connectionist approaches.) In such a network, a multilayered network of nodes encodes patterns of behaviour. Nodes on one part of the network (e.g., the bottom) would link, for example, to percepts, while higher-level nodes represent more abstract notions. For example, a low-level node would be able to read a colour value and a higher-level node or subnetwork of nodes would be able to interpret that as red, where red is a linguistic concept. Early approaches to this have worked for toy examples (see Harnad (1990); Hinton et al. (1986)). More recently, neural networks have used more principled approaches on how the nodes function, how the networks are constructed, and how they learn from data.

properties (or features) and aspects of language. It is through this grounded learning that such a model can learn the "meaning" of visual words, such as colour and shape terms. This is presently explained for definite descriptions and demonstratives in greater detail.

Grounding and Definite Descriptions

Definite descriptions that refer to visually present objects presumably are made up of words that help a listener distinguish an object from other objects by expressing properties that the referred object has. Thus words denoting colour, shape, size, spatial placement, etc., are used. The meaning of these words are not abstract (compared to, for example, the word *love*) where the meaning, it seems, is directly related to how those concepts are perceived visually. For example, if one attempts to explain the meaning of the word *red* to an individual who has never visually experienced red, it would be very difficult without pointing to objects that are red.[18] Rather, a word with a grounded meaning representation is based on perceptual information.

In the RE *the red circle*, the meaning of the word *red* is more of an ability to determine the redness of an object, given visual features of that object. In other words, given the features of an object, a meaning representation of *red* would be able to tell if that object is red or not. Likewise, the meaning of the word *circle* is the ability to determine something's roundness; the entire RE gives both words a vote as to how well they fit a particular object–that it is red enough and round enough to be distinguished from other objects.

This can be learned from data: given enough REs containing the word *red* and the objects to which those REs referred, a range of features values can be observed as being described as red with some potential uncertainty. The model of RR would need to somehow select the features and the range of values that denote what is referred to as being red.

Grounding and Demonstratives, Pronouns

Are words such as *it* and *that* grounded as descriptive words are? Pronouns refer to linguistic antecedents, but the exophoric kind we are interested in can also be grounded in a similar way as descriptive words. Recall our formal representation for pronouns and demonstratives, repeated in (19):

(19) a. $\iota x.recently\_referred(x)$
b. $\iota x.pointed\_at(x)$

[18] Red is, of course, a primary colour and cannot be derived from other colours. However, even though a colour that can be described in terms of others, such as purple which falls directly between red and blue, it still isn't quite sufficient to explain the meaning of purple as being a mix of red and blue.

Here grounding would be the same as it was with descriptive words: learn a function, either $recently\_referred$ or $pointed\_at$, and map between those concepts to the respective objects that fit into those classes. That is, learn a function that picks out the objects that were either recently referred or are currently being pointed at, respectively. Such a function for pronouns is fairly straight-forward: an object that is referred receives some kind of property that it was the last one that is referred, though pronouns aren't necessarily grounded in visual information, rather in contextual discourse information (though the referent is visually present). In contrast, grounding demonstratives is based on visual features, namely, knowing to which object the pointing gesture is indicating.

2.4.2 Grounding, Semantics, and Probabilities

We now look at how words grounded in perception fit into our formal framework. Following Equation 2.19 above, instead of representing an object x an abstract identifier, we can represent the object directly as some kind of feature vector, where the features represent visual aspects of that object. Equation 2.20 shows a slight modification to Equation 2.19, where x is now the vector $\mathbf{x}$:

$$[\![w]\!]_{obj} = \lambda\mathbf{x}.\phi_w(\mathbf{x}) \tag{2.20}$$

That is, we can represent a predicate ϕ_w using some kind of function that *learns* (i.e., not pre-defined) a mapping between object features and a decision that the features of the object are a good fit to what the concept represents. For example, a grounded function for red, i.e., $\lambda\mathbf{x}.red(\mathbf{x})$, would return 1 if the features x that represent some object are deemed by that function to sufficiently represent the concept of redness. In fact, instead of returning 1 if the object is deemed red enough and 0 if it is not, the grounded function for red can return a score, e.g., a probability between 0 and 1, where the closer to 1 the score is, the more red it is deemed to be. This is a type of probabilistic/stochastic approach to learning the grounded functions of words. Such a learning can happen by using example data of interactions, e.g., observing referring expressions and the features of the objects to which they refer, where the grounding function learns what features distinguish objects from being a good fit to that word and those features that do not.

The difference between a symbolic approach to meaning and a grounded approach to meaning is illustrated in Figure 2.7. For the symbolic approach, the world is represented as a set of classes (i.e., properties, both work in this example) where a particular object is denoted via a rule as being a member of a particular class by the judgement of a human designer. Often, this is done by a direct mapping between the word and the name of the class (e.g., some kind of orthographic distance function). Meaning in the grounded approach requires that a function between a word and a class be learned through data; this is illustrated in that there is a line between every word and every class, where the thickness of the line represents the degree as to how much that particular word belongs to that particular class. Furthermore, the class names are arbitrary as long as they are consistent.

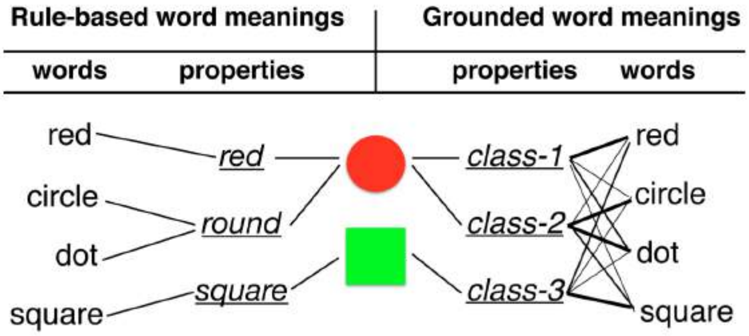

Figure 2.7: Symbolic, rule-based word meaning vs. a grounded word meaning. The rule-based approach requires words and class names to be defined, the grounded approach learns these mappings from data.

The models presented in Chapters 5 and 6 are grounded models. In the former model, a mapping between words and pre-defined properties is learned. In the latter model, the need for properties is eliminated and words are mapped to low-level object features. Both models are probabilistic and learn from example data.

Overcoming the Shortcomings

Recall the three shortcomings with FOL explained above

1. the set of classes must be determined

2. the functions that assign objects to those classes must be determined

3. incremental composition

When coupled with grounded semantics, these shortcomings aren't quite as constrained as they were before. We can represent objects in various ways, for example using low-level features or a set of defined properties. Using stochastic methods, we can use machine learning techniques to learn the functions that map words and concepts in FOL. This is explained in greater detail in later chapters. A stochastic approach also allows us to relax the uniqueness constraint in that a probability distribution over all objects is produced instead of a singleton set as a result of a computed truth value. The final constraint, that of bottom-up composition, has not been addressed through our discussion of grounding, but is of high importance, as we saw in Section 2.1. Incremental considerations are taken into account within the framework of each model and explained in their respective chapters.

The models presented in Chapters 5 and 6 link in their own way to FOL and perform a kind of grounding. Each also resolves REs in an update-incremental fashion, addressing the issue of bottom-up composition. The models are explained and evaluated in their respective chapters.

2.5 Reference and Meaning

Grounded semantics and FOL have both been used as approaches to approximating meanings of words, expressions, and sentences. The purpose of this section is to provide an overview of some philosophical ideas of meaning as it pertains to reference. Though not the direct focus of this thesis, the meanings of words and expressions are necessarily at issue when grounded semantics meets a semantic formalism like FOL, as we are attempting to do here. This section provides us with some additional terminology, in particular that of *intension* and *extension* which we use throughout this thesis.[19]

2.5.1 Philosophical Background

To begin, consider the following:

(20) a. ●
b. the red circle

At first glance, it may seem that the red circle represented in (20-a) is in fact the *meaning* of the RE found in (20-b). This could be generalised into a simple definition of meaning; that the meaning of a RE is, in fact, the object to which it refers. This, of course, doesn't really capture meaning at all, neither for the RE, nor for the words that make up the RE. The difference between the referent of a RE and the meaning of a RE can be distinguished more clearly in an example given in Chapter 1, repeated here:

[19] A more in-depth analysis of the philosophy of reference and meaning has been handled elsewhere; particularly as it pertains to proper names and indexicals, which we are not concerned with in this thesis. The reader is referred to Kellerwessel (1995) and Abbott (2010) for a more in-depth analysis of theories of reference (at the writing of this thesis the former, originally written in German, does not have an English translation). A more psychologically-motivated, yet philosophical stance on meaning and reference can be found in Burge (2010).

(21) The morning star is the evening star.

Where *the morning star* and *the evening star* are two different REs (albeit not the attention-directing type we are interested in) with distinct meanings, but they both in fact refer to the same object, namely the planet Venus. This example was explained by Gottlob Frege (1892) ("Über Sinn und Bedeutung" / "On sense and reference"), where, among other things, he explained the difference between *sense*, what we would call the meaning or notion of a word or expression, and the *reference*–that is, the referred entity itself. The entity itself isn't the meaning, rather it instantiates something to which the sense can refer. This distinction is made clear by another example using a triangle, with three additional lines, each connecting the midpoint of each line on the triangle with the opposite corner. These three lines inside the triangle intersect at a single point. The intersection of any two of those lines is the reference, but the two lines themselves that intersect at that point are the sense. Thus there are two senses, but one reference.

Before Frege, Mill (1846) made a similar distinction between *connotation*, the properties or attributes that are implied by a word or expression, and *denotation*, what an expression applies to in the world (i.e., the referred entity). This distinction is illustrated in Example (2) in Section 2.2, where in (2-a) no particular dog is being referred; the usage of the word *dog* connotes a type of entity that has properties belonging to dogs, and (2-b) where K is denoting a particular dog (which has all of the properties that the word *dog* connotes).

Russell (1905) asserted that properties (or, in his words, universals) are meanings of words (which, according to Abbott (2010) is getting at a Fregean sense). Definite descriptions, such as REs, aren't constituents, but rather quantificational NPs. Some definite descriptions are denoting descriptions, though they don't necessarily denote actual things, for example:

(22) The king of France is bald.

has a meaning, but does not denote an entity in the actual world because, at the writing of Russell's paper, there was no king of France–yet the definite description is clearly understood. Russell has a somewhat different view, then, on meaning when compared to Frege and Mill.[20]

A more pragmatic view was taken by Strawson (1950) claiming that there is a difference between the usage of a definite description and the definite description itself. Put another way, definite descriptions don't refer to things, *people* refer to things when they use definite descriptions. He further asserted that, when a definite description is used (preceded by the word

[20] Masheb et al. (2011) was quite critical of Russell's theory, distinguishing between denotation (i.e., the class of stuff denoted by a connotation) and reference, which is an individual or set instantiation of an entity of a particular class (e.g., *red* denotes red things, while reference refers specifically to an object that is red; i.e., in the class of red things).

the) the listener presupposes that the object exists, and that there is only one entity that is being referred (as explained above). Kripke (1977) also distinguished between speaker reference and semantic reference in that though the two are related, they perform different tasks.

Another important distinction can be made between REs that actually refer to real entities, and REs that are *attributive*. From Donnellan (1966):

(23) Smith's murderer is insane.

The example in (23) can be used attributively if it follows an utterance such as *anyone who would kill Smith must be crazy*, without particular care as to the individual who did the murdering, the point of the RE was to assign attributes to Smith's murderer–whoever it may be. Contrast that to *which one killed Smith?* to which comes the reply in (23), which helps the person who asked the question pick out the person who is the murderer.

Discussion These distinctions between sense/connotation, reference/denotation; speaker vs. semantic reference; attributive vs. attentional reference are things to consider when modelling a dialogue system component that resolves references to visually present objects, particularly using the above FOL as a modelling framework. Abbott (2010) notes that the semantics of attributive and referential REs are identical, at least in terms of a semantic formalism like FOL. In a dialogue setting, the speaker that utters a RE is certainly making a speaker reference to an object, but the listener must take the words of the RE, and (via semantic reference) resolve the object which the speaker referred.

Here, we follow Strawson (1950) that the speaker performs a RE in order to (also following Quine (1980)) draw the listener's attention to a particular instantiation of an object that fits the description of the RE. These are pragmatic considerations when references are being made. However, here we make the explicit assumptions that these are in fact taking place–all other types of reference (attributive, or non-referring) are not considered here.

Another focus of philosophers of meaning and is propositional truth values. For example, that *the King of France* has no referent and therefore the truth value must be false. But if one were to say, *John believes that the King of France is bald* must necessarily be false because *the King of France* has no truth value, but the fact that John believes it could very well be true (see Footnote 22). In this thesis, focus is not put on truth values of propositions. It is assumed that a RE does in fact refer and is therefore true. In terms of FOL, it is assumed that there is an object that satisfies the formula, and that there is only one (i.e., existence and uniqueness). Holding these things constant, we return again to sense/connotation and reference/denotation distinction.

2.5.2 Intension and Extension

As we are interested in resolving REs to visually present objects, it seems intuitive that we simply focus on the latter and forget the former. This would work if every object had a unique name that referred particularly, without ambiguity, to that object (and the philosophers mentioned in this section had a lot to say about names as they pertain to meaning and reference). However, when dealing with definite descriptions, such as *the red circle*, the referent is determined by the constituents of the RE, namely the words and their composition (another term from Frege where the meaning of sentences–or utterances–are composed by the meanings of their constituent parts).

Both Frege and Mill make a distinction between the meaning of a RE and the *designation*, or the thing to which a RE can be applied. Another, more technical term that is used to represent the notion of meaning (i.e., sense/connotation) is *intension* and the term used for that which is designated (i.e., reference/denotation) is *extension*. Indeed, the FOL framework that we opted for above is an intensional system of semantics in that it attempts to represent meaning by intension rather than extension.

Rudolph Carnap (1988) explained intension and extension in the following way in order to explain his semantic system using *properties* and *classes* (the connections between this and our above FOL explanation will become clear).[21] A property is something an entity has, whereas a class is something to which an entity belongs. The extension of something is akin to the class to which that something belongs; the intension of something is the corresponding property which that something has. Carnap gives the following examples (p. 18):

(24) a. The class Human is the same as the class Featherless Biped.
b. The property Human is not the same as the property Featherless Biped.
c. The property Human is the same as the property Rational Animal.

That is, Humans and Featherless Bipeds are extensionally equivalent. But they are not intensionally equivalent (i.e., they don't have the same Fregean sense). Carnap then continued to provide a semantics in which model-theoretic entities are identified with intension. This semantic system has been influential on later work in formal semantics, forming a group of *intensional logic* like our chosen FOL above.[22]

[21] He admits and later shows that the terms 'properties' and 'classes' aren't completely necessary in order to describe his semantic system, but he spans several pages using those terms as scaffolding to help the reader understand what he means by intension and extension, then discards them (as it were).

[22] Another technical explanation of intension uses the notion of *possible worlds*. Abbott (2010) explains this idea in the following way (p. 53-54):

> The value of incorporating possible worlds into one's semantics is that we can recognise a reference

For the formula in 2.20, ϕ_w is the intension, whereas the functional application of $\mathbf{x}$ to that intension forms the extension–i.e., the degree of belief that $\mathbf{x}$ fits ϕ_w. The question now is how to learn the intension, but the ideas presented thus far don't give us any idea of how to do so. The remainder of this section addresses this.

2.5.3 Intension via Extension

Kathleen Dahlgren (1976) makes the following claim in criticism of semantic theory (i.e., intensional logic):

> Semantic theory for natural language is faced with the following problem. It is relatively straightforward to formally state the semantic properties of whole sentences (such as ambiguity), and the relations between words and sentences (such as "not S is the negation of S"). It is even possible to give formal accounts which seem to be somewhat accurate, for how the meaningful parts, that is, the morphemes, phrases, and constituents of sentences combine to produce meaningful generative syntax or logic, some semantic properties of human language can be described. But no explanatory account of the semantics of individual words has been achieved using such methods. The theory of the lexicon has proven a difficult

> or extension for expressions not only in the actual world but also in other possible worlds. The possible worlds formulation of the notion of the INTENSION of an expression brings these extensions together, and give us something like a Fregean sense. In the most straightforward system, intensions are uniformly functions from possible worlds to extensions.

The idea behind possible worlds is that there is a set of (an infinite) possible alternatives to the way things are. Certainly, the world (i.e., the universe) is the way it is. The words we use to describe objects are what they are, e.g., the word *red* has a sense of 'red-ness'. But we could imagine another world where everything is exactly the same as the one in which we currently live, except instead we use the word *derf* to refer to what we know in this world as 'red-ness'. Words aren't the only thing that could be different in an alternative possible world; historical events might have turned out differently, e.g., Julius Caesar might have avoided his death on the Ides of March, or Columbus might have gotten completely lost and not ended up in what we now call the Americas. It is important in language partly because we can entertain concepts and ideas that are not necessarily ground truth in the world that we perceive. For example, I can say something quite absurd such as *I believe that everyone has the same favourite colour*. This is certainly not the case in the real world, but in a different world, i.e., a possible world that I have constructed in my mind, this statement could very well be the case.

The intension of a concept (i.e., a word, term, or expression), then, is a function from all possible worlds, where that concept applies, to the extensions. For our purposes, the intension is a function that picks out of each possible world whichever visually present object fits that description–in that particular world.

Though a seemingly elegant explanation of intension that takes propositional attitudes into account, it isn't completely necessary to hold this view for our models of RR to visually present objects. Resolving an extension using a characteristic function doesn't require us to entertain the notion of possible worlds.

> subject for philosophers and linguists alike.

This is more or less another way of putting the shortcomings to FOL we listed above. She continues (p.7):

> The meaningfulness of language lies in the fact that it is about the world.

Which follows from the above section on grounded semantics. She then makes the claim that extensionalism is preferred over intensionalism because, (following Putnam (1973, 1975)), natural language is not the property of individuals, but rather, is a social tool for communication. This focus on meaning in society was also taken up by DeVault et al. (2006) in that the meaning of a word (or, for our purposes, RE) is agreed upon by linguistic communities (see also Section 2.4 on grounding) by their usage, and not by individual speakers. Yet, somehow individuals need to be able to use language with other members of a language community and must somehow have some kind of "mental" approximation of what a word or RE means. In other words, language and meaning is established on societal level, but meanings do need to be represented in individuals somehow (i.e., meaning is to some degree approximated in the head of an individual (Chomsky, 1986; Pietroski, 2003; Daniels, 1990)). It is this interaction between individual and language community where, for the purposes of this thesis, intension should be explored.

Thus (continuing with Dahlgren (1976) p.14),

> Extensions determine intensions, though in a complex way, and not the other way around.

In other words, we can learn ϕ_w (i.e., the intension) by exposure to extensions and REs that have ϕ_w as a concept. To illustrate, suppose, for example, two friends, A and B, are walking down the street together. A is the member of a particular linguistic community (e.g., that of Chinese speakers), and B is not. As they walk down the street, A points to an object and utters something that B has never heard before. B can perceive what is being pointed at, in this case a stationary object, and get an idea of what A meant by the utterance. They continue walking down the street and A points to another object and says the same thing. This continues for some time and B begins to notice that all of the objects which A pointed to had very different features of shape, size, spatial placement, distance from them, etc. However, they all had the same colour, namely what B would call *red*. Later, B points to a red object an utters the word that A had uttered, with positive feedback (e.g., nodding) from A that tells B that the word was used correctly. One might say that, through interacting with someone who is a member of a linguistic community and perceiving objects (i.e., extensions), B was able to learn the word that picks out the concept for redness (i.e., the intension) in that language community, via reference. As noted in Chapter 1, the same is the case for children learning their first language (Wittek and Tomasello, 2005).

This example illustrates, at least conceptually, what Dahlgren claimed; namely that it is through exposure to extensions that intensions can be learned (i.e., approximated as to how they fit with a linguistic community). The intension, then, becomes the mechanism that assigns real-world entities to classes–something that FOL does not address (as explained in Section 2.3), but something which is essential in a practical dialogue system component that resolves references made to visually present objects. We show in Chapters 5 and 6 (but more particularly the latter), that intensions can be learned via examples of extensions, and that those learned intensions can be later used in RR tasks. It is through this procedure that we address the shortcomings of FOL using grounded semantics, thereby (following Daniels (1990), Larsson (2015), and Harnad (1990)) reconciling to a small degree symbolic and grounded semantics.

2.6 Additional Assumptions

Beyond the background given in this chapter, there are several additional assumptions that need to be established before moving on in this thesis. These assumptions allow us to focus on specific aspects such as grounded meaning and incrementally resolving REs.

Reference Perspective In this chapter, we have only seen examples where the speaker and the listener have the same *frame of reference*; i.e., their perspective on a particular scene is aligned as if standing next to each other. This, of course, isn't always the case when referring to objects. For example, a speaker that observes a scene with objects and knows that the listener is observing that scene from an opposite viewpoint (e.g., from across a table), the speaker might attempt to refer to an object based on the listener's perspective (e.g., as part of the RE, saying *on the left* would cause the listener to look to her left, which is the speaker's right). Work has been done in perspective alignment, e.g., Steels and Loetzsch (2009); Liu et al. (2010), but for this thesis it is assumed that the speaker and the listener have the same perspective of the scene.

Saliency When someone looks at a set of objects and determines to refer to one of those objects, what is it about that object, when compared to the others, that makes that person pick out certain features? For example, if there are 5 objects in the room and they all have the same colour, then using the referred object's colour won't distinguish that object from the others. Objects often have features that distinguish them from others, even features that stand out (i.e., are more *salient*) from other features. Knowing what features stand out from others could be useful in determining which object is being referred. However, as it is not directly related to REs, we are not interested in this thesis with measuring saliency directly.[23] For recent work on using saliency information to determine visually present landmarks in a navigation task, see Götze and Boye (2013).

Objects Defined What is an object? In this thesis, an object is a visually present entity that is distinct from other entities. In general, these objects also have distinct and unique properties, such as colour (i.e., a single a colour and not made up of multiple colours), shape, spatial arrangement, and size. We are not concerned with how the real world is perceptualised in in a human's brain; whether objects are represented symbolically or as a visual abstraction. We of necessity represent objects (in order to ground language with the world), and how an object is represented does have implications on how grounding takes place. However, in this thesis an object is clearly visible and distinct from other objects and has some kind of representation of which our models make use.

Universe of Discourse / Reference Domain The *universe of discourse* (Boole, 2005) in this thesis is the set of candidate objects that could be referred; i.e., all of the visually present objects that exist in a speaker and listener's immediate surrounding. This is always a small set of objects. It is always assumed that it is one of these objects that is being referred in a given RE, and not some other object that is unseen or was referred to at a different time. The task of resolving references means choosing which object from this set is the referred one. In other terminology, we assume a *reference domain* (Salmon-Alt and Romary, 2001) which are theoretical constructs that are entities which are presupposed at each use of a RE. The reference domain in this thesis is the set of visually present objects (we look into this a little bit more in Chapter 4).

Pragmatic Assumptions There are many speech acts, such as greeting, requesting, confirming, etc. (Searle, 1976). In this thesis, we are only interested in a single speech act, that of RE, such that a listener's attention is directed to an object. We also assume that grounding (i.e., establishing common ground) as in Clark (1996) where mutual understanding between the two participants has taken place to an extent, that the speaker and the listener have learned the meaning of the words used in the REs, but as we won't be looking at REs over the course of a discourse or dialogue, i.e., we mostly look at REs in isolation; that there isn't really an ongoing establishment of common ground. Moreover, we are only interested in two dialogue partici-

[23] This isn't completely true. In Chapter 5, we show in one experiment we take contextual saliency into account–something which that particular model can easily incorporate.

pants: a speaker and a listener (sometimes called the *addressee*); as explained, the models of RR that we explain take the place of the listener.

Other Assumptions As noted earlier in this chapter, we are only interested in referring to a single object (per RE). Certainly, a RE can refer to one, two, or a group of objects, which has been looked into in other work (Sauppé and Mutlu, 2014; Gorniak and Roy, 2004), but here we leave it to future work to handle REs that go beyond a single object. However, the models that are explained in this thesis should easily be extendible to be able to handle reference to multiple objects. It is also common for reference to be made to entities that don't exist in the immediate surroundings, but do certainly exist (e.g., New York, the moon, or a speech that was given). The models presented here could be extended (as we show, one of the models can refer to abstract things, if represented sufficiently) to refer to non-visual entities, but we are primarily concerned with those that are visually present. Also important, is that the objects are visually present at the moment the RE is taking place, and not at some other point in time. The objects can be perceived visually at the same time as the incrementally-unfolding RE event.

2.7 Chapter Summary

To couch the work in this thesis in an area where the models could be implemented as practical components, Section 2.1 explained *spoken dialogue systems*, particularly situated and incremental spoken dialogue systems. The types of referring expressions that this thesis focuses on were explained in the following section, namely definite descriptions, demonstratives, and exophoric pronouns.

Section 2.3 focused on what the listener does when resolving a referring expression and used a well-established formalism of first-order logic to do so. Examples of resolving an object using first-order logic for the three types of referring expressions were given. Some shortcomings were explained, namely

1. the set of classes must be determined
2. the functions that assign objects to those classes must be determined
3. incremental composition

and Section 2.4 addressed some of those shortcomings by appealing to *grounded* semantics.

Section 2.5 looked at philosophical literature on meaning as it pertains to reference and established a small aspect where our models make a contribution to theories of meaning. In short, the meanings of words and expressions are agreed upon by language communities, but individuals need to be able to approximate those meanings in their own heads in order to use language with other members of that community. Those meanings–intensions–are learned through interactions with real-world objects–extensions–and referring expressions that designate those objects. Learning the concepts and the mechanisms for determining class membership of an object is a contribution of this thesis.

The final section noted some additional assumptions that must be made for us to focus on the aspects of reference that this thesis addresses.

The overall goal is to model the resolution of referring expressions to visually present objects sufficiently that the model can be implemented in a practical component that would be usable in a spoken dialogue system that interacts with a human. Such a system would need to learn some notion of meaning of the words that are used as referring expressions, and it would need to be able to combine those word meanings in a practical way. Moreover, such a component would need to be situated and (update-) incremental because of the nature of the task which requires that objects be visually present.