Societal Grounding is Essential to Meaningful Language Use

David DeVault
[1]Department of Computer Science
Rutgers University
Picataway, NJ 08845-8020
David.DeVault@rutgers.edu

Iris Oved
Department of Philosophy
Rutgers University
New Brunswick, NJ 08901-1411
irisoved@eden.rutgers.edu

Matthew Stone[1,2]
[2]Human Communication Research Centre
University of Edinburgh
Edinburgh EH8 9LW, UK
Matthew.Stone@rutgers.edu

Abstract

Language engineers often point to tight connections between their systems' linguistic representations and accumulated sensor data as a sign that their systems really mean what they say. While we believe such connections are an important piece in the puzzle of meaning, we argue that perceptual grounding alone does not suffice to explain the specific, stable meanings human speakers attribute to each other. Instead, human attributions of meaning depend on a process of *societal grounding* by which individual language speakers coordinate their perceptual experience and linguistic usage with other members of their linguistic communities. For system builders, this suggests that implementing a strategy of societal grounding would justify the attribution of *bona fide* linguistic meaning to a system even if it had little perceptual experience and only modest perceptual accuracy. We illustrate the importance and role of societal grounding using an implemented dialogue system that collaboratively identifies visual objects with human users.

Introduction

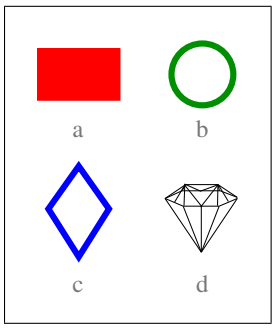
In this paper, we treat meaningful language use as an explicit design goal for conversational systems: they should mean what they say. We argue that achieving this goal requires that implementations explicitly connect system meaning to societal standards. Part of using language meaningfully, we claim, is working to keep your meaning aligned with what others mean in your community. Our case rests on strong intuitions about speaker meaning in specific contexts. We draw on these intuitions to understand and to criticize implementations that construe meaning in exclusively perceptual terms, and to articulate an alternative approach.

Systems that use language face a traditional line of objection that any meaning in a computer program's "utterances" is merely parasitic on the intentions and interpretations of its programmers. In its strongest form (Searle 1980), the problem is seen as endemic to computation: *none* of the symbols in a computer program, including any linguistic representations it may have, are ever intrinsically meaningful *to the system*; at best, it is argued, we engineer and arrange them in such a way that they *seem* meaningful *to us*. Other well-known arguments dispute the meaningfulness of language use in specific extant systems; the symbols they use to achieve linguistic meaning have been variously held to be hopelessly impoverished due to the limited range of inferences the system is able to draw using them (Dreyfus 1979), limited problematically to the "narrow micro-world" of the programmer's chosen domain theory (Winograd & Flores 1986), or effectively meaningless due to the lack of any coupling with the external world via perception (Harnad 1990).

The general thrust of these objections is that people find it hard to ascribe genuine meaning, of any sort, to disembodied, decontextualized, perceptionless computer programs. A common response has been to see the key to meaning as lying in a process of *perceptually grounding* a computer program's representations in real sensor data (Harnad 1990). While originally formulated as a strategy for imbuing arbitrary internal symbols with meaning, this approach has been thought to apply straightforwardly to symbols that link words to the world; indeed, many AI researchers explicitly advocate achieving linguistic meaning through perceptual grounding (Oates, Schmill, & Cohen 2000; Roy & Pentland 2002; Cohen *et al.* 2002; Yu & Ballard 2004; Steels & Belpaeme 2005).

This paper contests this view. Perceptual grounding, as it has been understood, is neither necessary nor sufficient to justify the attribution of *linguistic* meaning. Human interlocutors achieve stable, specific meanings in linguistic communication by coordinating their perceptual experience with linguistic usage across their community through a process we call *societal grounding*. Participating in this process is a prerequisite for intuitive ascriptions of speaker meaning, and yet realizing this process robustly would justify saying a system meant what it said even if it had little perceptual experience and only modest perceptual accuracy.

Our argument links questions about meaning in implemented systems to the phenomenon of meaning borrowing described in the philosophy of language (Kripke 1972; Putnam 1975). We use this connection to develop a systematic characterization of societal grounding, and then sketch an architecture for realizing societal grounding within the information-state approach to dialogue management (Larsson & Traum 2000). Our concluding discussions position computation as a framework that can continue to inform the study of meaningfulness in machines and in humans.



(U_1)	C:	*it's a square*
(U_2)	U:	*you mean a rectangle?*
(U_3)	C:	*no*
(U_4)	U:	*which square?*
(U_5)	C:	*the square*
(U_6)	U:	*there is no square*
	C:	[ignores this]
(U_7)	C:	*the square is solid*

Figure 1: User interaction with the COREF agent. The user (U:) can see the four displayed objects, but not COREF's (C:) private labels {a,b,c,d} for them. The target in this example is object a.

A motivating example

Figure 1 shows an excerpt of an interaction with COREF, an implemented dialogue agent designed to collaboratively identify visual objects with human users (DeVault *et al.* 2005). In this interaction, COREF's goal is to get the user to identify object a, the solid red rectangle. COREF begins by uttering U_1, *it's a square* — even though the target object is not a square. In order to diagnose the remainder of this dialogue, and in pursuit of our goal of meaningful language use, we would like to understand what COREF means by its uses of the term 'square' in U_1, U_5, and U_7. If it does not mean what its human users mean when they say 'square', we would like to understand why not.

As it happens, COREF represents the meaning of its uses of 'square' with an internal symbol, square. COREF classifies any object x as square when

$$\begin{array}{cc} & (i) \quad \texttt{rectangle}(x) \\ \& & (ii) \quad \frac{\texttt{length}(x)}{\texttt{width}(x)} \leq 1+\varepsilon \end{array} \tag{1}$$

In this case, COREF classifies object a as square, because a is classified under COREF's rectangle symbol, and its dimensions satisfy (ii). Any perceptual classifier would need a tolerance threshold such as ε to allow for noise in its length estimates. The exact value ε takes is unimportant; we caricature COREF's misclassification in our depiction of a.

We will use this interaction to gradually develop a fine-grained understanding of the interplay between linguistic meaning, mental meaning, and system design. For notation, we will write **square** for the property that COREF's user means by 'square', for example in utterances U_4 and U_6. The questions we now wish to examine carefully are: Does COREF mean **square** when it says 'square'? If not, is this because COREF's internal symbol square does not mean **square**? What *engineering* would be required in order to make COREF mean **square**?

Linguistic meaning and speaker's beliefs

To make our argument for the role of societal grounding in linguistic meaning, we will examine a series of alternative conditions that might be thought necessary in order for an agent A to mean M by a use of linguistic term T. Each condition can be seen as an attempt to specify under what circumstances A could be said to "understand" what is allegedly meant. For example, according to the historically influential "description theory of meaning", competent speakers know, for each term, some description or set of properties that isolates the term's meaning; see e.g. (Devitt & Sterelny 1999). We might endorse this theory with the following condition:

> A does not mean M by T unless A associates T's meaning with a set of properties that uniquely identifies M. (2)

The properties that COREF associates with the meaning of 'square' in (1) do not uniquely distinguish instances of **square** (as evidenced by object a), so the description theory counsels that COREF does not mean **square** by 'square'.

However, in the 1970s, compelling arguments emerged in the philosophy literature that, contrary to the description theory, a speaker's meaning in using a term depends not just on what the speaker privately believes about the term's meaning but also strongly depends on perceptual and social factors (Kripke 1972; Putnam 1975; Burge 1979). From the standpoint of AI, the most important arguments against the description theory are what philosophers have called the problems of *ignorance* and *error* (Devitt & Sterelny 1999): people mean what they say even when they lack complete and correct knowledge of meaning. For example, Putnam (1975) argues persuasively that a human speaker can use the words 'elm' and 'beech' to mean **elm** and **beech** (the two types of tree) despite being unaware of *any* (non-linguistic) property that distinguishes elm trees from beech trees. When such a speaker asks of a tree "Is that an elm?", he still inquires meaningfully whether it is an instance of **elm**. Nor do erroneous beliefs seem to undermine linguistic meaning; a human speaker can still mean **elm** when he says "Is that an elm?" even if he falsely believes that every tree to which 'elm' applies is less than 30 meters tall.

In designing a meaningful agent, we take it as a methodological constraint that we ought not impugn the meaningfulness of the system on grounds that could also impugn the meaningfulness of its human users. Thus, (2) is not a *general* principle according to which we can deny the meaning **square** to COREF's utterances of 'square'. The inaccurate properties in (1) that COREF associates with 'square' *could* just be analogous to the properties a meaningful human speaker erroneously associates with 'elm'.[1]

Nevertheless, something is clearly wrong with COREF: it doesn't seem to mean **square**, and this fact seems to explain the miscommunication in Figure 1. Philosophers have suggested that the mechanism by which a human speaker is able to mean **elm** when he says 'elm', despite being unable to characterize that meaning accurately, is that there is a *division of linguistic labor* within a community (Putnam

[1] Parallel possibilities of ignorance and error rule out (2) as a condition on human meaning for mass terms like 'gold' or 'water' (Kripke 1972; Putnam 1975), medical terms like 'arthritis' (Burge 1979), artifact terms like 'pencil' (Putnam 1975) or 'chair' or 'sofa' (Devitt & Sterelny 1999), or for names like 'John' or 'Socrates' (Kripke 1972), among others.

1975). The idea is that some members of the community of human English speakers have had perceptual contact with elms, and thereby have acquired an accurate fix on that type of tree; other community members who are less experienced are somehow able to "borrow" the meaning of the term 'elm' from these "experts". A non-expert speaker's meaning is thus determined not by his private beliefs about or experience with elms but instead by a chain of "meaning borrowing" events that begins with the speaker, leads along some path through the members of the community, and finally arrives at an expert whose experience does fix the meaning (Kripke 1972).

To date, philosophers have not produced a clear theory of the fine-grained perceptual and social details needed to fully explain this "meaning borrowing" mechanism; see (Devitt & Sterelny 1999) for a recent survey. In the meantime, however, AI researchers have made considerable progress in articulating the sorts of detailed connections between system representations and sensor data that might justify attributing genuine meaning to implemented agents' *mental* states. In the next two sections, we explore, in the context of this research, how perceptual experience and social interactions might combine to produce genuine *linguistic* meaning in implemented systems.

Linguistic meaning and speaker's perception

Perception is the gateway between an agent's mental state and the external world, and as such it clearly plays an important role in assigning external meanings to an agent's internal representations. One tempting way to formalize this role is to simply *identify* a representation's meaning with some perceptual measure. For example, Oates, Schmill & Cohen (2000) identify the meaning of the verb 'pushed' with distinctive sequences of robot sensor values. Gorniak & Roy (2004) identify the meaning of 'green' with a probability distribution over image RGB values. Cohen *et al.* (2002) define the meaning of 'square' as the probability that 'square' is uttered given that an object perceived as having a certain height-to-width ratio is under discussion.

Understanding linguistic meaning in such direct perceptual terms might lead us to endorse the following "discrimination condition" on linguistic meaning:

> A does not mean M by T unless A can perceptually discriminate instances of T's meaning, M, with high accuracy. (3)

According to the discrimination condition, whether COREF means **square** when it says 'square' depends on how accurately it can perceptually discriminate instances of **square**. Suppose COREF were able to infer the properties in (1) from sensor data and thereby achieve a perceptual accuracy of, say, 90%.[2] Would this perceptual accuracy be sufficient for COREF to mean **square**?

The problem with requiring any particular accuracy as a requirement for linguistic meaning is again that we will *inevitably* deny linguistic meaning to human speakers who we strongly believe have it. For example, an "ignorant" human speaker who asks of a nearby tree, "Is that an elm?", and is privately unaware of any distinguishing difference between elm and beech trees, may very well lack the capacity to perceptually distinguish elms from beeches. Yet even lacking this perceptual skill, he can still meaningfully ask whether a certain tree is an instance of **elm**. Thus, the discrimination condition suffers from a problem of *perceptual ignorance*, and it therefore cannot serve as a general principle according to which we could deny a meaning like **square** to an agent based solely on its modest perceptual accuracy.[3]

Despite the possibility of limited perceptual accuracy, perhaps a speaker A's perceptual experience with instances of M is still the ultimate arbiter of A's meaning. Perhaps A has to *at least* have perceived *some* instance(s) of M in order to linguistically mean M by some term T. For example, Yu & Ballard (2004) use eye tracking and image segmentation techniques on camera images to build a perceptually grounded representation of the object (a piece of paper) that a human user is fixating when she utters the word 'paper'. Perhaps such a historic perceptual exposure must ground an agent's internal representation of what a term means before linguistic meaning is possible:

> A does not mean M by T unless A represents T's meaning by some internal symbol S that means M because A has perceptually grounded S in instances of M. (4)

According to this "perceptual exposure condition", COREF could not mean **square** unless its internal symbol `square` were grounded by historic perceptual exposure to squares.[4] We believe there is a substantive "causal insight" underlying this condition, which brings several advantages to a theory of linguistic meaning. First, identifying linguistic meaning with the external cause of a perceptual experience (Kripke 1972) liberates the bearer of that meaning from necessarily having accurate belief or perceptual discrimination, which we have already found problematic.[5] Second, it suggests a "causal-historical" analysis in which every meaningful mental symbol can be traced back to some historical perceptual event that gave it its meaning. If such an event can be described as connecting a mental symbol to the external object or property that is its meaning using only the causal vocabulary of physics (Harnad 1990), then mental meaning will have been "naturalized". Further, condition (4) suggests that linguistic meaning can be reduced to such mental meaning, so that these two difficult problems can be solved at once.

[2] In fact, COREF currently works from a hand-built domain representation that includes the properties in (1). I.e. it does not infer these properties from sensor data.

[3] The possibility of perceptual ignorance also rules out (3) as a condition on human meaning for the terms in footnote 1.

[4] In fact, COREF's symbol `square` is not so grounded. COREF's "perception" currently begins with a hand-built domain representation rather than real sensor data.

[5] The fact that A has causally interacted with some instance of M (e.g. by way of a camera) does not entail that A's beliefs about M will be accurate, or that A will be able to recognize other instances of M when A sees them.

But it must be possible for the connection between linguistic meaning and perceptual experience to be less direct than condition (4) suggests. Our "ignorant" human speaker who inquires, "Is that an elm?" may well have never seen an instance of **elm** before.[6] Yet we strongly believe he *still* means **elm**. Somehow, this perceptually inexperienced speaker seems able to borrow this meaning after mere *linguistic* contact with other speakers who use the *word* 'elm' (Kripke 1972). This ability of a human speaker to speak meaningfully about things he has not directly encountered does not mean that speakers' perceptual histories are irrelevant to their linguistic meanings. However, it does mean that we cannot appeal to (4) as a general principle to deny linguistic meaning to an implemented agent like COREF.

Linguistic meaning and societal grounding

In the previous two sections, we have argued that linguistic meaning is not precluded by various inadequacies in a speaker's beliefs, perceptual abilities, or perceptual experience. An ignorant and perceptually inexperienced speaker can still achieve meaning through meaning borrowing and the division of linguistic labor. This doesn't mean meaning is easy to achieve. On the contrary, we think meaning borrowing itself must be understood as a hard-fought achievement. Language affords us the crucial freedom to speak to one other about things some of us may not have first-hand experience or expertise with, but that freedom carries special responsibilities to coordinate with others about what we believe and what we perceive. It requires us to *work* to keep our private representations of linguistic meaning aligned. Thus, even in cases of meaning borrowing, these responsibilities continue to make belief and perception essential to linguistic meaning.

We call an agent's coordination of its linguistic meanings with the community *societal grounding*. Building on the causal insight underlying condition (4), we can describe societal grounding as a relation between a speaker, an internal symbol, and a linguistic meaning:

Definition.[7] Agent A has *societally grounded* internal symbol S in meaning M iff both conditions D1 and D2 are met:

D1. S means M, because either:

1. A has perceptually grounded S in instances of M, or
2. A has heard some other member B of the linguistic community use term T, where:

(a) In fact, B meant M by T, and
(b) A uses symbol S to represent what B meant by T.[8]

D2. A is committed to an active and consistent cognitive role for S. In particular:

1. A is *linguistically committed* to S:

(a) A entertains S as a representation for what other speakers mean by their terms in conversation, and
(b) When A derives a representation involving S to capture the content of some proposition P that has been (credibly) asserted or implied in conversation, A reconciles[9] P with private beliefs and perceptual classifiers that use S.

2. A is *perceptually committed* to S:

(a) A entertains S as a representation for what A perceives when A acquires a perceptual experience, and
(b) When A derives a representation involving S to capture the content of some perceptual experience E, A reconciles[9] E with private beliefs and perceptual classifiers that use S.

We express a speaker's obligation to societally ground his representations of linguistic meaning as follows:

> A does not mean M by T unless A represents T's meaning by some internal symbol S that A has societally grounded in M. (5)

Condition (5) and clause D1(2)(a) thus make societal grounding and linguistic meaning interdependent.

In order for a speaker A to mean **elm** by 'elm', according to (5), A must have societally grounded some internal symbol, say `elm`, in the meaning **elm**. Requirement D1 distinguishes two alternative origins for the connection between `elm` and its meaning. If A perceived an instance of **elm** and thereby perceptually grounded the symbol `elm`, then D1(1) is satisfied. Alternatively, D1(2) allows that A may instead be using the symbol `elm` as a representation for whatever some *other* agent, B, meant by some term.[10] If what B meant was **elm**, A "borrows" **elm** as the meaning for `elm`.

Requirement D2 demands, as the price of meaning **elm** in conversation—by way of symbol `elm`—that the speaker remain "engaged" with new evidence from other speakers and from perception that may bear on what `elm` means. If A uses `elm` to represent A's meaning in asking a local expert "Is that an elm?" about a nearby tree, D2 obliges A to try to exploit the answer to improve A's beliefs and perceptual skill at recognizing instances of whatever it is that `elm` means.

Concretely, A's belief state might include statistical information like $P(\texttt{height}(x) \leq 30\text{m} | \texttt{elm}(x)) = 0.9343$, and A might perceptually classify a perceived object x under its `elm` symbol using a set of camera image features $I(x)$. We can view these aspects of A's mental state as constituting A's "understanding" of `elm` as a representation of linguistic meaning. Since `elm` really means **elm**, we can say that A understands *more* about what its symbol `elm` means when

[6] Let's suppose the tree in question is a beech not an elm.

[7] This definition is intended to characterize a natural phenomenon that we believe exists and deserves further investigation.

[8] To be clear: We do not intend the meaning of S to be identified with the *description* "what B meant by T", but rather with the referent of that description. The symbol S means M, the external-world individual or property that is in fact B's meaning, whether or not A has any independent experience or understanding of this individual or property. What A knows for certain about what S means is that the causal connection between S's meaning and S is mediated by B. Note also that there is no guarantee that A does not have a different symbol that, unbeknownst to A, also means M.

[9] as A's other goals permit

[10] Most likely, the term B used was 'elm'.

A's belief state involving `elm` is closer to being *true* and when *A*'s perceptual classifier for `elm` is more accurate.

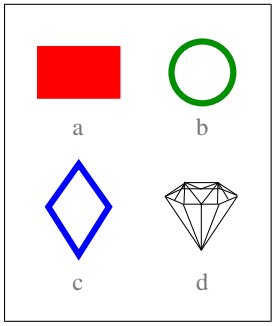
If *A* derives the representation `elm(tree3)` to represent the local expert's answer of "Yes" to *A*'s question about the nearby tree, requirement D2(1)(b) obliges *A* to adopt some strategy to reconcile `elm(tree3)` with *A*'s understanding of what `elm` means, to keep `elm` societally grounded. For example, $P(\texttt{height}(x) \leq 30\texttt{m} | \texttt{elm}(x))$ might be re-estimated using the observation $\texttt{height}(\texttt{tree3}) = 32.7\texttt{m}$, and *A*'s perceptual classifier might be re-trained using a positive instance with image features $I(\texttt{tree3})$.

D2 might appear to be just a basic condition on rational "meaning borrowing". If *A* created the symbol `elm` with the intention that it would mean whatever some other agent *B* meant on a certain occasion, that intention imposes a commitment for *A* to persist in tracking what *B* meant by exploiting new evidence as it arrives; cf. (Cohen & Levesque 1990). Yet something deeper is at play here: even if *A* *perceptually* grounded `elm` using *A*'s own prior experience with instances of **elm**, *A* may still understand very little about what kind of things in the world `elm` has been linked up with. Due to a poverty of experience, *A*'s belief state and perceptual classifiers for `elm`, which encode *A*'s understanding of what else would be the same kind of tree, may be wildly or uncomfortably inaccurate. This is why requirement D2 applies even if the origin of the meaning of *A*'s symbol lies in *A*'s own perceptual experience.

Thus, on the view we are advocating, meaning is always seen by an agent as an external *target* that the agent needs to think about and act in pursuit of. This gives meaning an important place in an agent's mental life and in psychological theory, as summarized by requirement D2, even though the facts about meaning, as indicated by requirement D1, are always determined by factors external to the agent. We believe it is the commitment that language speakers have to this pursuit, rather than the truth of their evolving beliefs or the accuracy of their evolving perceptual skills, that underwrites our ascriptions of linguistic meaning to each other.

For example, we began by noting how very hard it seems to attribute the meaning **square** to COREF's utterances of 'square' in the interaction of Figure 1. We are now in a position to state why this is, and to answer the questions that motivated our various attempts to characterize linguistic meaning clearly. The problem is *not* that the properties in (1) which COREF associates with what it means by 'square' do not isolate the property **square** (though they do not). The problem is *not* that COREF lacks a perceptual classifier that discriminates instances of **square** with high accuracy (though it lacks one). Nor is the problem that COREF has not perceptually grounded its symbol `square` in sensory experience with instances of **square** (though it has not).

The problem is that COREF just isn't committed in the ways it needs to be to mean **square** when it speaks. At utterance U_6, the user categorically denies that **square** is instantiated in the visual display. According to requirement D2 for societal grounding, COREF ought to recognize that the user's assertion conflicts with its belief state for its `square` symbol and *do* something about it. But as currently implemented, COREF ignores any user utterance it cannot interpret as a clarification question or as a move in its collaborative reference task. This leads COREF to ignore this utterance, along with the evidence it could have provided about the symbol `square`. Alas, COREF proceeds to call object `a` 'square' once more in U_7. This lack of commitment to its symbol `square` is why COREF does not mean **square** when it says 'square'.[11]



(U_1)	C:	*it's a square*
(U_2)	U:	*you mean a rectangle?*
...		
(U_6)	U:	*there is no square*
(U_7^*)	C:	*isn't the solid one a square?*
(U_8^*)	U:	*no*
(U_9^*)	C:	*ok*
(U_{10}^*)	C:	*it's not a square*
(U_{11}^*)	C:	*it's the solid one*

Figure 2: Hypothetical user interaction with an improved COREF agent. Utterances U_1-U_6 are as in Figure 1. Utterances U_i^* are hypothetical and do not reflect COREF as currently implemented. In this interaction, COREF uses U_7^*-U_8^* to societally ground its symbol `square`.

Implementing linguistic meaning

We have argued that it takes special *work*, over and above what is needed for *mental* meaning, for an agent to achieve *linguistic* meaning. To illustrate how we can begin to implement this extra work, Figure 2 shows a hypothetical interaction with an improved version of COREF that implements societal grounding. Let's suppose that COREF has previously borrowed the meaning **square** for its symbol `square` from one of its users, and that COREF currently classifies objects as `square` using the properties in (1). According to D2(1)(a), COREF should entertain `square` as representing what the user means by 'square' in U_6, "there is no square". Since `square` *does* represent the user's meaning, COREF should derive a representation like $\neg\texttt{square}(\texttt{a})$ as implied by U_6. Further, COREF should try to reconcile $\neg\texttt{square}(\texttt{a})$ with its private beliefs and perceptual classifiers. Since (1) classifies `a` as `square`, COREF's strategy for implementing its commitment might lead it to confirm the user's implication by asking U_7^*, "isn't the solid one a square?" When the user confirms in U_8^* that $\neg\texttt{square}(\texttt{a})$, COREF could discharge its commitment by some algorithmic adjustment to the factor ε in (1).

In fact, we can begin to implement societal grounding as achieved in interactions like this one using techniques that already allow dialogue systems like COREF to reason flexibly about meaning and the evolving state of a collaborative conversation. Figure 3 illustrates how COREF could represent this particular negotiation. As the conversation pro-

[11] It is very hard to state what COREF's symbol `square` *does* mean, if anything, since COREF does not currently implement perceptual grounding or meaning borrowing.

Information State: Task Stack	Information State: Facts	Utterance	Interpretation
CollabRef(t)	square(a), solid(a)	(U_1) C: *it's a square*	C asserts addcr(t, square(t))
CollabRef(t)	square(a), solid(a), square(t)	(U_2) U: *you mean a rectangle?*	U does Downdate(U_1) U does Push(CollabRef(S)) U does Push(YNQ) U asserts ynq(addcr(S, S = rectangle))
YNQ CollabRef(S) CollabRef(t)	square(a), solid(a)	(U_3) C: *no*	C asserts no(addcr(S, S = rectangle))
CollabRef(S) CollabRef(t)	square(a), solid(a)	(U_4) U: *which square?*	U does Pop(CollabRef(S)) U does Reinstate(U_1) U does Push(WhichQ) U asserts whichq(addcr(t = _))
WhichQ CollabRef(t)	square(a), solid(a), square(t)	(U_5) C: *the square*	C asserts addcr(t, t = a)
CollabRef(t)	square(a), solid(a), square(t), t = a	(U_6) U: *there is no square*	U does Downdate(U_5) * U does Push(NegotiateFacts) U asserts addFacts(¬square(_))
NegotiateFacts CollabRef(t)	* ¬square(a), solid(a), square(t)	(U_7^*) C: *isn't the solid one a square?*	* C does Push(YNQ) C asserts ynq(addFact(square(a)))
YNQ NegotiateFacts CollabRef(t)	* ¬square(a), solid(a), square(t)	(U_8^*) U: *no*	* U asserts no(addFact(square(a)))
NegotiateFacts CollabRef(t)	* ¬square(a), solid(a), square(t)	(U_9^*) C: *ok*	* C does Pop(NegotiateFacts)
CollabRef(t)	* ¬square(a), solid(a), square(t)	(U_{10}^*) C: *it's not a square*	* C asserts addcr(t, ¬square(t))
CollabRef(t)	* ¬square(a), solid(a), ¬square(t)	(U_{11}^*) C: *it's the solid one*	* C asserts addcr(t, t = a)
CollabRef(t)	* ¬square(a), solid(a), ¬square(t), t = a		

Figure 3: A trace of the information states and utterance interpretations that COREF could employ to model the user interaction of Figure 2. Representations for utterances U_1-U_5 reflect the implemented COREF agent, while representations marked by an asterisk (*), for utterances U_6-U_{11}^*, are hypothetical and do not reflect COREF as currently implemented.

gresses, COREF maintains an evolving context representation called an *information state* (Larsson & Traum 2000) that tracks the tasks and subtasks that are underway, the facts that are agreed by the interlocutors, and other contextual details. In particular, COREF's information state already represents and tracks the intended speaker meanings from prior utterances in order to interpret and answer clarification questions like U_2, as in (Purver 2004).

COREF generates and understands utterances by drawing on grammar and context to link words to actions that transform the information state. For example, Figure 3 shows how U_1 corresponds to the action addcr(t, square(t)), which adds the fact square(t) to the information state; this fact serves as a constraint on the the target t which COREF intends the user to identify. The clarification U_2 temporarily cancels or "downdates" the effects of U_1 on the information state and starts a nested collaborative reference subtask to identify the value of S, a variable introduced into the context to represent COREF's meaning in using the word 'square' in U_1. COREF's task knowledge specifies a network of possible next actions for each task, including the possibility of opening or closing nested subtasks as in U_2. COREF also approaches each task with a strategy for selecting actions to further the task and achieve a favorable outcome. These representations already give COREF the ability to interpret and answer yes–no and wh-questions like U_2, U_4, and U_7^*.

Accordingly, Figure 3 suggests that modeling societal grounding in COREF is a matter of extending COREF's existing repertoire with suitable additional grammatical resources and collaborative activities. We treat U_6 as a

move that starts one of these new activities: it pushes a `NegotiateFacts` subtask in which interlocutors will work together to resolve conflicting representations of aspects of the current situation. In this case, the issue is which objects are squares. The subsequent question–answer pair applies COREF's existing language and task representations; the effect is that COREF agrees that there are no squares, pops the negotiation subtask, and brings its revised understanding back to the ongoing process of collaborative reference.

Of course, we'll need more than just this representation and architecture to implement systems that pursue such dialogues effectively. One obvious challenge comes in scaling up models of grammar and collaboration to the range of human linguistic behavior that naturally occurs in meaning negotiation dialogues, like $U_6 - U_9^*$ in Figure 2. Another is developing strategies to accommodate conflicting evidence about linguistic meanings, as when different users use the same term in different ways. Are there two different meanings, or is one of the users wrong? A system could confront users over any unexpected usage, or simply follow what they do (while aggregating noisy evidence into a single model of linguistic meaning). More generally, we will need strategies to balance meaningfulness with other system design goals.

The theory of societal grounding suggests that responses to these challenges will fit together to help us build open-ended agents that work with their interlocutors to extend their competence. Of course, robust systems would see wide deployment, in areas from information management to entertainment, wherever it proves infeasible to circumscribe users' linguistic behavior during system development. Designing for open ended linguistic competence seems a mystery, or even an impossibility, to authors like Winograd and Flores (1986). To us, it is a matter of correctly interfacing social, linguistic and perceptual abilities, and when broken down this way it seems amenable to familiar methodologies, particularly in language technology.

Meanings, communities and languages

Our approach connects speakers to the external-world objects or properties they mean in conversation. It does this by distinguishing having a *representation* of something external from having accurate *understanding* of that thing. Understanding tends to come slowly, with experience, over time, but a speaker can have a meaningful representation by simply listening and forming the right commitments. Thinking computationally about this distinction can help us clarify the cognitive and social processes required for meaning in humans as well as machines. We will give three examples: characterizing our deference to experts, describing how meaning connects language users to each other, and linking our shared meanings to our individual knowledge of language. This discussion may clarify our view as it pertains both to models of practical dialogue and to the general project of understanding intelligence as computation.[12]

[12]While this paper addresses only perceptible, external-world meanings—due to our interest in dialogue systems that deal in such meanings—we believe the presence of societal grounding may help explain conversation about entities that no one has ever perceived.

According to the division of linguistic labor, meaning grounds out when a chain of meaning borrowing terminates in an "expert". This expert is a speaker who has perceptual experience of some object or property, who draws on this perception (rather than what others mean) for his own linguistic meaning, and who manifests sufficient understanding that other speakers choose to borrow that meaning. In terms of our definition of societal grounding, experts have perceptually grounded some representation of the external world (as in D1(1)), are committed to that representation (as in D2), choose to represent their own linguistic meaning in terms of that representation (as in (5)), and serve as sources for other speakers who elect to borrow that meaning from them (as in D1(2)). Clearly, any speaker can become an expert for any meaning. Expert status just requires that the expert and other community members choose to ground their linguistic meanings in the expert's perceptual experience rather than someone else's.

The interconnectedness between speakers who borrow meanings from each other can make it tempting to think of linguistic meaning as "distributed" throughout a network of language users, rather than located in any particular place. The fact that an "ignorant" English speaker means **elm** when he says 'elm', rather than some other meaning, depends on a chain of historic meaning borrowing events in which different community members each play their part at a specific point in the past. By analogy, the fact that this speaker means **elm** is similar to the fact that the speaker has a certain person *P* as their great-grandmother: what makes *P* the speaker's great-grandmother is a sequence of historic events in which different people played important parts at specific points in the past. While we wouldn't ordinarily say that the *fact* that *P* is the speaker's great-grandmother is "distributed" throughout the historic members of the speaker's genetic community, we might say that the *explanation* for this fact, or the set of factors that make it true, is so distributed. Likewise, societal grounding provides a historic, community-based *explanation* for the fact that an "ignorant" English speaker means **elm**, but it does not suggest that the fact itself, or the meaning itself, is somehow "distributed".

With this perspective, we can now see that societal grounding remains compatible with a fairly traditional understanding of linguistic knowledge (Chomsky 1995). Cognitive science is concerned with the psychological mechanisms that govern the internal linguistic representations in a speaker, or what Chomsky calls I-language (internal language). Yet each speaker also has their own idiosyncratic cognitive representations of the external world (as they have experienced and understand it), including their representations of linguistic meaning. Where we depart from Chomsky is in narrating the cognitive mechanisms of societal grounding not only as internally realized computations but also in terms of explicit connections between speakers and their external environments. If our view is correct, then speakers themselves pursue these connections, not necessarily through a single, shared abstraction, like "the English language", but rather through their own perception and a network of local authorities whose expertise provides a path to meaning.

Conclusion

We conclude with two contributions our perspective brings to the methodology of system building. First, our approach shows how conversational systems can borrow, represent, and *use* meanings from the start, and *learn* about them while they use them. Thus we can narrate the interaction in Figure 2 as COREF learning something about what speakers mean by 'square', while at the same time using the word 'square' with a stable, consistent, and correct meaning. Such a description would not be possible if linguistic meaning were identified with a perceptual measure. For example, even if COREF were to retrain its `square` classifier after the user expresses $\neg$`square(a)`, COREF would use 'square' with a *new* meaning in U^*_{10}, because the perceptual classifier that *is* the agent's meaning would have changed.

Second, by allowing us to ascribe public meanings to what our systems say, we can be clear about what they mean and whether they are right or wrong. This way, we can try to build them to be right. It might be worried that societal usage can never fully settle the correctness of an agent's usage, because different speakers can mean different things by the same terms. But such ambiguities are fully compatible with societal grounding as the arbiter of what is really meant: condition D1 establishes a *correct* meaning, for each mental symbol, according to the actual chain of events that connects a symbol to its meaning. Of course, agents must still select particular symbols to represent what other speakers mean; societal grounding explains how ambiguities can arise without putting the facts about meaning up for grabs.

Contrast this with the multi-agent simulation of Steels and Belpaeme (2005), in which agents maintain weighted associations between linguistic terms and different perceptual classifiers that serve as alternative linguistic meanings for the agents. The agents realize a commitment to communicative success in society by adjusting these weights to encourage success and penalize failure in a simple language game. But in the absence of clear standards for correctness in linguistic meaning, the model lumps together all failures in communication, even though, for example, being misunderstood and right seems very different from being understood and wrong—as one strongly suspects COREF is in utterance U_1 of Figure 2.

In the end, we believe that bringing our semantic intuitions into correspondence with system design is essential for building agents we can relate to. To act meaningfully, conversational agents must faithfully realize the interrelated capacities that underpin our judgments of meaning in interactions with one another. We have argued that societal grounding is such a capacity, so that as implemented techniques for societal grounding become more robust, users will become more comfortable attributing *bona fide* linguistic meaning to implemented systems; conversely, AI researchers will be able to present more principled evidence that their systems really mean what they say.

Acknowledgments

This work was supported by the Leverhulme Trust, Rutgers University and NSF HLC 0808121. We thank our anonymous reviewers, Mike Harnish, Alex Lascarides, John Pollock, Mark Steedman, and Anders Strand.

References

Burge, T. 1979. Individualism and the mental. In French, P.; Uehling, T.; and Wettstein, H., eds., *Midwest Studies in Philosophy: Studies in Metaphysics*. Minneapolis: University of Minnesota Press. 73–122.

Chomsky, N. 1995. Language and nature. *Mind* 101:1–61.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Cohen, P. R.; Oates, T.; Beal, C. R.; and Adams, N. 2002. Contentful mental states for robot baby. In *Proceedings of AAAI*, 126–131.

DeVault, D.; Kariaeva, N.; Kothari, A.; Oved, I.; and Stone, M. 2005. An information-state approach to collaborative reference. In *ACL 2005 Proceedings Companion Volume. Interactive Poster and Demonstration Sessions*, 1–4.

Devitt, M., and Sterelny, K. 1999. *Language and Reality*. Cambridge, MA: MIT Press.

Dreyfus, H. L. 1979. *What Computers Can't Do*. Harper and Row.

Gorniak, P. J., and Roy, D. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21:429–470.

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

Kripke, S. 1972. *Naming and Necessity*. Harvard.

Larsson, S., and Traum, D. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering* 6:323–340.

Oates, T.; Schmill, M. D.; and Cohen, P. R. 2000. Toward natural language interfaces for robotic agents. In *Proceedings of Agents*, 227–228.

Purver, M. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. Dissertation, University of London.

Putnam, H. 1975. *Mind, Language and Reality*. Cambridge. chapter The Meaning of 'Meaning', 215–271.

Roy, D., and Pentland, A. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science* 26(1):113–146.

Searle, J. R. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3(3):417–457.

Steels, L., and Belpaeme, T. 2005. Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences* 28(4):469–529.

Winograd, T., and Flores, F. 1986. *Understanding computers and cognition: A new foundation for design*. Addison-Wesley.

Yu, C., and Ballard, D. H. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception* 1:57–80.