Use theories of meaning

between conventions and social norms

Marc Staudacher

Use theories of meaning

between conventions and social norms

ILLC Dissertation Series DS-2010-07

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Use theories of meaning
between conventions and social norms

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op donderdag 2 december 2010, te 12.00 uur

door

Marc Staudacher

geboren te Oakland, Verenigde Staten van Amerika.

Promotiecommissie

Promotor:
prof.dr. M.J.B. Stokhof

Co-promotores:
dr. P.J.E. Dekker
dr. R.A.M. van Rooij

Overige leden:
prof.dr. R.I. Bartsch
prof.dr. M.V.B.P.M. van Hees
prof.dr. R.G. Millikan
prof.dr. R.A. van der Sandt
prof.dr. F.J.M.M. Veltman

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

This research was supported by a Marie Curie Early Stage Research fellowship in the project GLoRiClass (MEST-CT-2005-020841).

Cover design by the author based on an illustration by John Tenniel.
Printed and bound by Ipskamp Drukkers B.V., Enschede.

ISBN: 978–90–5766–212–3

Contents

Acknowledgments

Hélicon: Et que voulais-tu?
Caligula, *toujours naturel*: La lune.
Hélicon: Quoi?
Caligula: Oui, je voulais la lune.

Caligula
Albert Camus

For the last three years, I've lived the life of a researcher with all the ups and downs that are part of it. I'm grateful to many persons and institutions that helped me with reaching this goal, even if it was ambitious (hence the epigraph).

I thank Nicholas Asher, Gerhard Jäger, and Hannes Rieser for supporting my application at the ILLC, and Thorsten Wilholt and Ansgar Beckermann for a timely confirmation of my degree.

In retrospect, getting the job was the easy part of this enterprise. I'm grateful to my three supervisors Paul Dekker, Robert van Rooij, and Martin Stokhof. They had the thankless job of supervising a student who at times would rather produce a new plan about his thesis than commit himself to work towards it. Thank you for your support and availability (especially in the final phase)! I learned a lot from you.

Special thanks go to Lars Dänzer, Michael Franke, Tikitu de Jager, and Peter Schulte. Lars was not only a good friend over the years, he also often told me what I really wanted to say (in a charming way) and commented on a late draft of my thesis which improved the result. Michael made my time in Amsterdam a pleasure. He also commented on my thesis but personally even more important, he introduced me to Yoga. That changed my life (for the better). I enjoyed the travels with Tikitu and our discussions. He proofread my whole thesis at an amazing pace (all remaining errors are mine). I enjoyed the reading groups with Peter. After that, his PhD thesis set me straight on normativity and metaphysics. Thank you friends! I learned a lot from you.

For their comments on my work I also thank Edgar Andrade, Max Kölbel, Ruth Millikan, Christoph Pfisterer, and Wolfgang Schwarz. Thank you, friends and colleagues! I also learned a lot from you.

I thank Ali Ben Horsting, Vincent Jansen, Diederik Klein, and Phuc Hoai Truong Nguyen for their help with the Dutch abstract.

I also have the discussions with the following friends and colleagues in good memory (and the ones I forgot apparently a little bit less): Dora Achourioti, Maria Aloni, Jochen Apel, Chantal Bax, Renate Bartsch, Anton Benz, Peter van Emde Boas, Benjamin Börschinger, Mirjam Bloemendaal, members of woongroep "Brandnetel," Ralf Busse, Mike Collins, Inés Créspo, Wayne Davis, Cédric Dégrémont, Michael Devitt, Alexander Dinges, Paul Egré, Andreas Erz, Dimitris Gakis, Amélie Gheerbrant, Kathrin Glüer, Thomas Grundmann, Fritz Hamm, Pelle Guldborg-Hansen, Frank Hofmann, Joachim Horvath, David Hunter, Simon Huttegger, Daisuke Ikegami, Matthias Koch, Sven Lauer, Johannes Marti, Nathalie Lötscher, Andy Lücking, Kirk Ludwig, Eric Pacuit, Peter Pagin, Christina Pawlowitsch, Werner Raub, Alexander Reutlinger, Francois Recanati, Lennart Ritterhoff, Floris Roelofsen, Hans Rott, Olivier Roy, Katia Saporiti, Friederike Schmitz, Silvère Schutkowski, Giacomo Sillari, Sarah Skupin, Brian Skyrms, Jakub Szymanik, Joel Uckelman, Noortje Venhuizen, César Sainz de Vicuna, Andreas Witzel, Lucian Zagan, Henk Zeevat, Jelle Zuidema, and Jonathan Zvesper. The regular reading groups "Philosophy of Language and Beyond" and "Classics in game theory" were fun. The feedback from participants of the Philosopher's Lunch and of the GAP Doktorandenworkshop was helpful.

For my research visits I'd like to thank Andreas Kemmerling (Heidelberg), Sarah-Jane Conrad and Klaus Petrus (Berne), and Hans-Johann Glock (Zurich).

I gratefully acknowledge the generous Marie Curie scholarship the EU awarded to me as a fellow in the GLoRiClass project Benedikt Löwe set up. I also thank the members of the Familienrat des Zyli'schen Legats for their financial support during critical phases in my PhD. Finally, I'd like to thank the ILLC for its excellent research environment; especially I thank

Karin Gigengack, Tanja Kassenaar, Peter van Ormondt, Ingrid van Loon, and Leen Torenvliet for their help.

My research has profited a lot from Open Source technologies. Notable are TeX(XeTeX, MikTeX), Notepad++, TeXNicCenter, and Subversion. On the closed software side, Citavi helped me a lot.

Personal thanks go to my family — Andy, Koni, and Veronika —, to my dear friends Nele Dechmann, Elisha Rüsch, Stefan Rusconi, and last but not least to Judith Raum. I love you. You gave me a lot of faith to get this done. Now, this is it. You don't have to read it.

Amsterdam Marc Staudacher
July, 2010.

Chapter 1

Introduction

> "When I use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean, neither more nor less."
> "The question is," said Alice, "whether you can make words mean so many different things."
> "The question is," said Humpty Dumpty, "which is to be master – that's all."
>
> *Through the looking glass*
> Lewis Carroll

Carroll's puzzling dialog illustrates the questions to which this thesis is devoted: *What makes words mean something? And why do they mean what they mean?* From this perspective, my interest in the argument between Humpty Dumpty and Alice does not concern whether Humpty Dumpty can make words mean *so many different things*, as Alice questions, but rather whether Humpty Dumpty can make words (phrases/sentences/...) mean something *at all.*

Humpty Dumpty's problem, as I like to understand it, is one about *what makes words of a public language mean what they mean in that language.* I do not want to understand it as a problem about what a *speaker* can mean and I also do not want to understand it as one about words which belong to an individual's *idiolect.* On the basis of this understanding, the problem is the following. Humpty Dumpty claims that he can make words mean what he wants them to mean. But it seems that he can't do that. It's not just Humpty Dumpty who can't do that. No single person can make a word mean something. Consequently, this holds for each person. Hence, no person can make a word mean something. This is counterintuitive: For who but us makes the words mean something and, in particular, mean what they actually mean? That is, something about Humpty Dumpty's claim must be right. So clearly, the above reasoning must be faulty.

The problem makes us aware of the fact that words mean something and that this is so because of certain other facts. Why should we care for these facts and why should we try to find out more about their relations? To a certain extent, I think that there is no point at all to pondering such questions. Language, after all, is a tool, and to the extent that it serves our purposes, there is simply no need to investigate it. We can leave it as it is, it seems.

However, even our everyday language use is full of questions concerning the meaning of our talk. We wonder what someone wanted to convey when she uttered the words she uttered. Were they a remark or rather a criticism? A question or a suggestion? Often, we wonder what someone meant when she said something (or didn't). We question whether someone really has said what he wanted to say. We are also interested in questions about "meaning" in a wider sense of *being about something* and *being informative about something.* We want to know where the words and their meanings come from. We're curious what their uses tell us about the persons using them. We think of some uses as being colloquial, of others that they are pretentious. Some indicate that the speaker belongs to the upper class and others that he doesn't. Sometimes, the use of a certain word is enough to figure out the region the speaker is from, and so on and so forth.

That is, even if we treat language as a tool, we want to know more about its inner workings. Questions about meaning belong here. Let me explain why. Language is so pervasive that it might seem that almost all our conscious activities involve it. If we think about it, then we realize that language is an incredibly complex form of social behavior. We can do many things by using language – like informing someone about something, asking someone a question, making a promise, or telling a story. Clearly, what we can do by using language depends – to a large extent – on what the words therefor used *mean.* A simple and rough *same-saying* test assures us of that: What I can do by uttering "I see a bachelor right now" is roughly the same as what I can do by uttering "I see an unmarried man right now." This is so because "bachelor" means the same as "unmarried man." And that's why I can't do the same by uttering "I see an unmarried woman right now." For "bachelor" does not mean "unmarried woman." In short, questions about meaning are at the center of the phenomenon of language use.

So meaning is important and a better understanding of it would be helpful. But Humpty Dumpty's problem shows that our understanding based on common sense is not good enough since seemingly contradicting opinions are part of it: (i) No one (of us) can make a word mean something in a public language and yet (ii) it is us who make them mean what they mean. Yet, we shouldn't abandon common sense too quickly.

A problem in the reasoning towards (i) was that only particular individuals were considered and not groups of them. However, linguistic communication is *social.* It typically involves at least two persons: a speaker and a hearer. – And to remind you, the way I want to understand Humpty Dumpty's problem is about languages used in *public* and not just by a Robinson Crusoe and his Friday who have their own language. Hence, the typical case is communication in a bigger social group and more precisely, *two person*-communication with a speaker and a hearer who are part of the group. – With this in mind, a rejoinder to the reasoning can be formulated as a question: *Can two persons make a word mean something?* The inclination seems to be a clear "no." But, there is something strange about this way of asking. For could three persons make a word mean something? The answer seems to be still "no" but how many persons are required, then? Hence, Humpty Dumpty's problem is better understood as providing a systematic answer to this question. To this end, we need a principled explanatory theory.

What should the theory be about? What should it explain? I understand these questions as a question about the kind of *project* one is engaging in when one tries to solve Humpty Dumpty's problem. I will call the project the "conventionalist project" since conventions play a central role in it. I will take up this point in the next section and provide here a short outlook about what is to come in this thesis.

Outlook In this thesis, I discuss the conventionalist project. I present and improve on theories from three different research paradigms, namely "Signaling Games,"[1] "Actual Language Relations," and "Evolutionary Theories." It turns out that none of the current theories has the conceptual resources to solve Humpty Dumpty's problem adequately. Hence a better theory is called for.

[1] I refer to research paradigms by capitalizing their names. I refer to the notion of a signaling game as defined in game theory by using small caps; likewise for the other paradigms.

According to the account I develop, it's true that it's our use and understanding of words (and other linguistic expressions) that determines their meanings. But not every use determines meaning. It's the use we convene on by means of *conventions* or the use that is regulated by *social norms*. The notion of a social norm is then what should be added to conventionalist theories to address Humpty Dumpty's problem. One of the differences between conventions and social norms, I'm going to claim, is that the latter are sensitive to social structure in an important way: Not every party to a social norm is on equal footing; some have power while others don't. In terms of power Humpty Dumpty's problem can be solved.

On my account meaning is determined in a rather different way than people claiming that "meaning is conventional" or that "use determines meaning" have thought. For social structure is usually not taken into account and conventionalists don't distinguish between conventions and social norms. This assessment is naturally sketchy but we will work towards a fuller understanding in due time.

1.1 The conventionalist project

The conventionalist project consists in providing an answer to the question "What makes the sentences of a semantic theory true?" Semantic theories entail sentences of the following kind:[2]

[2] This relates as follows to (i) Davidsonian theories of meaning and (ii) intensional semantics. Ad (i): My way of describing the semantic theories seems to rule out Davidsonian theories of meaning. But clearly it shouldn't. Traditionally, Davidsonian theories of meaning are understood as Tarskian truth theories. Hence, meaning sentences are not part of such a theory. But one can understand Davidsonian theories of meaning so that they contain at least one further axiom that justifies the transition from T-sentences to sentences of the form "e means m in $\mathcal{L}$" under certain conditions. This has been proposed by Kölbel (2001:618), following a suggestion by Larson and Segal (1995:560 fn. 15). On this understanding Davidsonian theories of meaning also contain meaning sentences (but only for sentences and not for sub-sentential expressions). Another way is simply to stick to the traditional understanding of Davidsonian theories of meaning and to modify the explanandum for such theories: A conventionalist theory for such a semantic theory has to explain what makes the sentences of a Tarskian truth-theory true. Ad (ii): Intensional semantics use abstract entities such as functions as meanings. But one shouldn't think that their meaning assignments do *not* have the form of meaning sentences. For I leave open what the entities are to which the names in the m-position refer. If an intensional semantics is used, they refer to their respective abstract entities.

(1) “Die Türe ist zu” means *that the door is closed* among German speakers.

(2) “Howdah” means *the reading seat on the back of an elephant*, currently in English among English speakers.

The examples illustrate a common pattern. They are of the form “expression e means m at $\mathcal{C}$.” An expression is mentioned in the e-position. A meaning is referred to in the m-position. In the *at-$\mathcal{C}$*-position is something serving as a *coordinate* The coordinate can be a language like English or be more complex, involving a social group, a time, a history, or a world.

We can learn different things about meaning. If you don’t know what “howdah” means, then you learn its meaning when someone tells you what it means. Yet, even if you’re perfectly informed about its meaning, you might be uninformed about the kind of facts in virtue of which “howdah” means what it does.

Conventionalists provide a theory about this second kind of information. Their question is one about the constitution of meaning of words and other sorts of linguistic expressions (phrases, clauses, sentences) in a public language. Hence, their question is not: “What does ‘howdah’ mean?”, but: “Why (in virtue of what) does ‘howdah’ mean what it means?”. Thus the question is a *meta-semantic* question and the goal is to provide a *foundational theory of meaning.* The distinction here is between semantic theories and foundational theories of meaning. The goal of the former is to provide a theory which entails for each sentence of a language a meaning sentence. The goal of the latter is to explain what makes meaning sentences true.[3]

Not any kind of foundational theory of meaning is accepted by conventionalists. They want to explain the meanings of expressions in terms of their *use.* Hence, conventionalists commit themselves to the following thesis:

U. For all expressions e, meanings m, coordinates $\mathcal{C}$: The use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

U leaves a lot to be desired: (i) What is meaning?, (ii) What is use?, (iii) What is it to determine an expression’s meaning?, and (iv) How might use do so? I’ll return to these questions below. The thing to observe here is that different accounts give different and sometimes no answers to these questions. But without making the thesis more precise one could object

[3] See (Speaks 2010) for a recent discussion of the distinction.

that conventionalists do not even hold a thesis with a determinate content.

However, we should not expect that all conventionalists answer these questions in the same way. There are alternatives and there is no reason to restrict the conventionalist project to a particular choice. Consequently, the task of answering these questions is at the level of individual accounts and not at the level of the project. But we can say at least the following about the use of expressions: it is a *social* one in a group and not a purely individualistic one. Consequently, conventionalist accounts are a kind of a (social) use theory of meaning. Their characteristic claim is that the use is *conventional*:

C. For all expressions e, meanings m, coordinates $\mathcal{C}$: The conventional use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

A further thesis which often comes up in debates about meaning is that meaning is normative:

N. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

I return to the N-thesis below in §1.1.3. So I'll be brief here. Like C, N is put in a noncommittal way. We'll have to refine them while working towards a serious account. The thesis is interesting for the conventionalist project since plausibly, *if meaning is normative, then because of the way it is determined – e.g.* because of certain "conventions." I return to this principle in chapter 2 where I also argue for a version of N.

The combination of C and N yields four basic positions:

$\mathbf{C_{-}N_{-}}$. Meaning is neither conventional nor normative.

$\mathbf{C_{-}N_{+}}$. Meaning is not conventional but normative.

$\mathbf{C_{+}N_{-}}$. Meaning is conventional but not normative.

$\mathbf{C_{+}N_{+}}$. Meaning is both conventional and normative.

A conventionalist claims that one of the last two positions is true. This is in need of an argument (see below in §1.3). But we can only start arguing when it has been made clear what the content of the claims is. This requires a clarification of the central notions of meaning (§1.1.1), (semantic) normativity (§1.1.3), and conventions and social norms (§1.2).

A conventionalist has two ways to proceed: She can start with the conviction that the conventionality thesis C and the – still to be stated – normativity thesis are analytic truths and stipulate the central notions in a way that the theses come out as analytic truths, no matter what it requires. This is the *stipulative* way.

The other way is the *explicative* way. It consists in using explications of the central notions which are independent of the project (and in particular, the respective notions of a convention and a social norm should be applicable to non-linguistic phenomena). The question is then: “Are the C/N-theses true under the used explications (or are there at least good reasons in favor or against them)?”. Importantly, the answers might be: “No, they are false under the most plausible explications of these notions.”

So, the interesting way seems to me to be the explicative one, especially since there is, *prima facie*, no reason against it. Hence, I’d like to understand the conventionalist project as one which relates meaning, conventions, and social norms as three independently characterized notions. If the project succeeds, an informative explanation is provided.

The plan is now as follows. In the remainder of this section, (i) I elaborate on the relevant notion of meaning, namely literal meaning. (ii) I explain how to understand the talk of “determination.” (iii) I provide a pre-theoretic characterization of (semantic) normativity. (iv) I provide an overview about current research paradigms within the conventionalist project.

Subsequently, I offer in §1.2 pre-theoretic characterizations of what is commonly called a “convention.” They are used subsequently in §1.3 to execute the argument that meaning is conventional. In §1.4 I introduce an adequacy condition for conventionalist accounts and return to Humpty Dumpty’s problem. In §1.5, the chapter ends with an overview about what is to come.

1.1.1 Literal meaning

Conventionalists claim that somehow, conventions determine the meanings of expressions. Let us focus on that what is determined by conventions: *meanings of expressions.* The conventionalist project, as I like to understand it, is about *literal meanings.* Hence, we should say more about what literal meaning is.

One could be tempted to say that literal meaning *is* conventional meaning.[4] But a conventionalist *shouldn't* claim this. For it would not only be uninformative but also trivialize the conventionality thesis C. Hence, another way of defining literal meaning is required. In a first step, I tentatively suggest a minimal characterization of a literal meaning. But even the minimal characterization has issues. In a second step, I describe a better way to define literal meaning.

It's common to distinguish different kinds of meanings, in particular (i) speaker-meaning and linguistic meaning and (ii) literal meaning and non-literal meanings. Literal meaning is a kind of linguistic meaning. Hence, the conventionalist project is about *literal linguistic meanings.* Linguistic meaning has at least the following two marks:

First, in contrast to speaker-meaning, linguistic meaning is a property of words and other kinds of linguistic *expressions* while speaker-meaning is a property of utterances.[5]

Second, sentences which have a linguistic meaning (typically) also have truth conditions (or rather *satisfaction conditions*) in context.[6] This is not to claim that everything about meaning consists in truth or satisfaction conditions. Grammatical mood (like indicative or imperative) are often also counted as an aspect of meaning, and so is social information (like social class) that is associated with an expression. I set these aspects aside and focus on the truth-conditional aspect.

A literal meaning is then something that is a linguistic meaning which has the following mark: It is akin to the dictionary meaning of a word, *e.g.* the literal meaning of "piglet" is *small pig.* Moreover, it contrasts with *non-literal meanings*, among them being the meanings of metaphors ("You are a piglet" in the sense of *You are a dirty child*),[7] hyperboles ("You are a genius" in the sense of *You are very smart*), and malapropisms ("I take for granite" in the sense of *I take for granted*).

There are issues with this minimal characterization of a literal (linguis-

[4] For example Recanati (2004:68) claims this.

[5] I use "utterance" in the wide Gricean sense of "utterance" including all sorts of tokening events, among them non-verbal behaviors and planted evidence.

[6] For a recent defense of the claim that meaning is truth conditional, see Lycan (2004). It's contested whether evaluative expressions like "Murder is wrong!" have truth conditions in a substantial sense. But at least in a minimal or deflationary sense such expressions can be said to be "true" or "false", *cf.* (Schulte 2008:§5.6).

[7] The example is from (Sperber and Wilson 1986:154).

tic) meaning of an expression. In particular, the property of expressions to have satisfaction conditions is contested (the second mark of linguistic meaning). For among those who loosely accept the minimal characterization,[8] there is disagreement about key properties. In the so-called "border war" in the semantics/pragmatics debate various notions of literal meaning have been produced that don't share a common core. Neo-Griceans – like Borg (2006) – claim that sentences should be assigned a meaning which is truth-evaluable. Others – like Sperber and Wilson (1995) – deny that sentences have such a meaning and assign them something which needs to be "enriched" in order to be truth-evaluable. This is to say that there is disagreement about how to understand the second mark of linguistic meaning and there is disagreement whether its preciser versions should be accepted or not. Moreover, the preciser versions have subtle consequences for the mark of literalness. If one allows for more a context-sensitive notion of a linguistic meaning (as Sperber and Wilson do), then more linguistic meanings are classified as literal meanings (and *vice versa* if one proposes less context-sensitivity). Given this disparity in key properties, it seems hopeless to find a common characterization.

So, we need to be more specific to make the talk of literal meanings precise but can't just easily factor out a common core. My proposal to fix the notion of literal meaning (for conventionalist purposes) is to use the following recipe:

The aspiring conventionalist goes to the pragmaticist of her choice who is in the possession of a theory of linguistic communication. Such a pragmatic theory explains minimally (i) what it is to mean something by uttering an expression, (ii) what it is to understand an utterance, (iii) what it is to successfully communicate, and other things like: what implicatures are and how they can be understood and so on.

The conventionalist asks the pragmaticist two things: (i) about her pragmatic theory and (ii) about her use of the term "literal meaning." The conventionalist uses the answers as follows. The first is used to define a mini-communication model. In terms of this model, the conventional use of expressions is described. On the basis of the description we can fix the use

[8] Davidson (1978:31 ff.) arguably rejects my characterization since he's not inclined to distinguish metaphorical meanings from literal meanings. He holds the thesis "that metaphors mean what the words, in their most literal interpretation, mean, and nothing more." But Davidson also rejects the conventionalist project (see chapter 3). So, I think we can put his position aside.

properties of expressions that enter the determination claim expressed by the conventionality thesis C. The answer to the second question is used to derive a role description for "literal meaning." Such a description captures the role of literal meaning in communication. Arguably this is not the only use of the notion but it is an important one. (Below I provide an example.)

This recipe makes sure that a conventionalist is doing the right thing, namely defining an interesting kind of meaning which is faithful to the minimal characterization and for which a foundational theory of meaning is provided. This is part and parcel of stating the conventionality thesis with a determinate content. Moreover, it creates harmony between her theory and the use of the so determined meanings in the pragmatic theory. For suppose that contrary to the recipe, the conventionalist account used a mini-communication model that is not coherent with the one used in the pragmatic theory of the overall framework.[9] – Such an approach is hard to evaluate in abstract. It seems to me to be very odd. Figure 1.1 illustrates the relevant relationships between the theories and characterizes a *framework for literal meaning.*

One popular theory-combination that fits this framework results from using a Gricean pragmatic theory. Such a theory explicates the use and understanding of expressions in terms of the Gricean notions (in some version) of speaker-meaning and -understanding. Successful communication is (roughly) analyzed as the hearer's recognizing the speaker's communicative intention. Non-literal meaning (implicatures, meanings of metaphors, ...) are explained in terms of general communicative principles and literal meanings. A conventionalist that is asking such a Gricean would define her mini-communication model accordingly by using the Gricean notions of speaker-meaning, understanding, and successful communication. A literal meaning of an expression is then whatever kind of meaning that can systematically play the roles in the Gricean pragmatic theory. One of these roles is to be something that explains together with the general communicative principles non-literal meanings. Another role is that it has an explanatory

[9] The mini-communication model and the pragmatic theory do not have to be *identical.* Arguably, mutual consistency in general is too weak. For two non-overlapping theories trivially satisfy this condition. Coherence seems to be right, which is roughly mutual consistency between theories that are conceptually connected. In our case, coherence between the mini-communication model and the pragmatic theory amounts to (i) be mutually consistent and to at least sharing the notions of (ii) meaning something by uttering an expression, (iii) understanding an utterance, and (iv) successful communication.

role in successful linguistic communication.

If one endorses this position, then one is committed to the underlying assumptions of the used theories. The prominent examples are the conventionalist accounts of Lewis (2002), Schiffer (1972), and Bennett (1976). One of the crucial commitments is that the first step of Grice's program – the step to analyze a basic notion of speaker-meaning in terms of propositional attitudes – can be carried out *without depending on expressions' already having a literal meaning.* (I'll argue against this position in §6.2.4.)

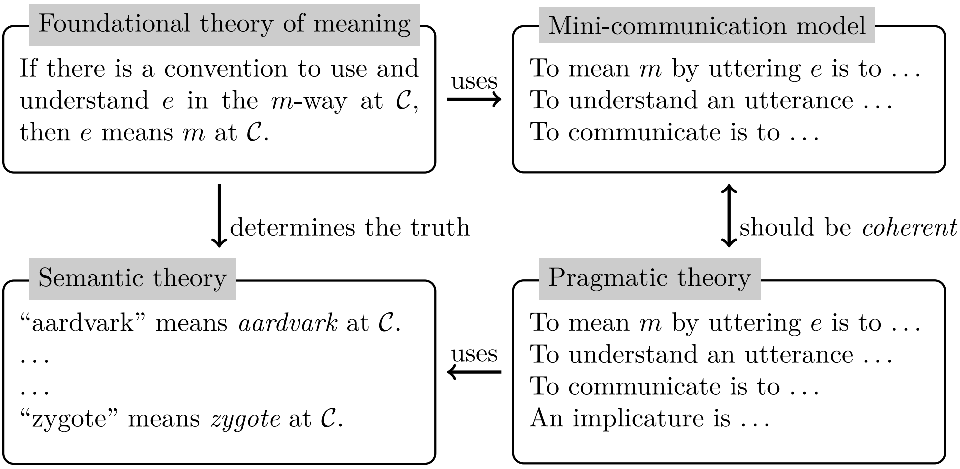

Foundational theory of meaning
If there is a convention to use and understand e in the m-way at $\mathcal{C}$, then e means m at $\mathcal{C}$.
uses
Mini-communication model
To mean m by uttering e is to ...
To understand an utterance ...
To communicate is to ...
determines the truth
should be *coherent*
Semantic theory
"aardvark" means *aardvark* at $\mathcal{C}$.
...
...
"zygote" means *zygote* at $\mathcal{C}$.
uses
Pragmatic theory
To mean m by uttering e is to ...
To understand an utterance ...
To communicate is to ...
An implicature is ...

Figure 1.1: A theoretical framework for literal meaning

1.1.2 Determination

The central thesis C of the conventionalist project – that conventional use of expressions determines their meanings – is stated in terms of *determination.* Sometimes, I'll also say that "there is meaning *in virtue of* conventions" to express the same thesis. Let me elaborate on what I mean by "determination". Generally, it's taken to be a cover term for different kinds of ontological dependency relations between properties or facts. Here, I propose to understand it in terms of global supervenience. Global supervenience has been studied in the debate about physicalism – the thesis that everything is physical. To make this thesis precise, it is usually stated as the thesis that everything supervenes on the physical. The latter thesis is commonly accepted as a necessary condition for the former; it's contested that the latter is also sufficient for it. The idea of supervenience is that there are different "levels" or "descriptions" of reality, *e.g.* the mental and the physical. Let me define the global-supervenience relation and then show how it connects to the conventionalist project.

Global supervenience is a binary relation that holds between sets of facts (or properties). For two sets of facts A and B, the A-facts *supervene globally on* the B-facts just in case different A-facts imply different B-facts. In a first approximation, we can define the relation in terms of *duplicates* of a world:

(3) A-facts supervene globally on the B-facts iff any two worlds that are B-duplicates are also A-duplicates.

More precisely, for some set of facts F, *being-an-F-duplicate* is an equivalence relation over the set of possible worlds: For any two possible worlds w and w', w' is an *F-duplicate of w* iff w' is with respect to F-facts indiscernible from w.

However, (3) has an unwelcome consequence: it's too strong. Its instances are of the form "A-facts supervene globally on B-facts," for some sets A and B. But they shouldn't be necessary claims but *contingent* claims about our world. For we don't want to rule out that there are *metaphysically* possible worlds that are inhabited by Cartesian souls or immaterial ghosts. We just want to rule out that our world is among these worlds:[10]

(4) A-facts supervene globally on B-facts iff any world which is a B-duplicate of our world is also an A-duplicate of our world.

But this is still not correct. It rules out that there are B-duplicates of our world that also contain a lot of A-like stuff that is sustained in non-B-like stuff, *e.g.* worlds that are physically just as ours but that are inhabited also by immaterial ghosts. But we don't want to rule out that such worlds are possible. We just want to make sure that a *minimal* B-duplicate of our world which duplicates just the B-facts *but nothing else* is also an A-duplicate.[11]

[10] See (Jackson 1998:11 ff.).

[11] *Cf.* (Jackson 1998:12). My definition of global supervenience is weaker than Jackson's since not all instances of GS do not entail that everything supervenes globally on the physical.

Hence, let us define global supervenience as follows:

GS. A-facts supervene globally on B-facts iff any world which is a minimal B-duplicate of our world is also an A-duplicate of our world.

Meaning determination in terms of global supervenience GS relates as follows to the conventionalist project: Let A be the set of meaning facts – facts of the sort that e means m at coordinate $\mathcal{C}$ – and let B be the set of facts that there are conventions at these coordinates concerning the use of the expressions mentioned in the meaning sentences. Instantiating GS in this way amounts to the thesis that *meaning supervenes globally on conventions.* This seems a plausible way of explicating the conventionality thesis C. So, I suggest to understand the conventionalist project as endorsing GS.

Yet, this instance of GS alone does not capture the conventionalists' determination-claim. On the one hand, GS only entails that the B-facts covary with the A-facts and not that the A-facts *depend* on the B-facts.[12] On the other hand, GS does not explain why the covariation holds (or, in other words, what the *nature* of the covariation is).

So, GS is not sufficient to capture the essential claim of conventionalists. In the debate on physicalism, people have defended stronger theses; among them is the conceptual-entailment proposal by Frank Jackson (1998) which I find plausible.[13]

The conceptual-entailment proposal consists in the claim that there is a conceptual connection between the A- and the B-facts: the set of B-propositions conceptually entails A-propositions. This is to say that, at least in principle, the A-propositions are *a priori* derivable from the B-propositions.[14]

Jackson's proposal is elegant since it cashes out the desired dependence between the A- and the B-facts and also explains why the B-facts covary

[12] See (Kim 1998:9–13) for a helpful discussion.

[13] For a defense of it, see also (Chalmers and Jackson 2001). Schulte (2010) offers a helpful clarification with respect to reductive explanations from this perspective.

[14] (i) GS is stated in terms of *facts*; entailment is a relation between *propositions.* So, we have to relate them. We can do so, by requiring that the facts are described in a canonical way in some favorite vocabulary, *cf.* Schulte (2010) on *basic descriptions*: Let p be a basic description of a fact, then *the fact that p* corresponds to *the proposition that p.* (The use of basic descriptions is required for other reasons, *cf.* (Schulte 2010).) (ii) Conceptual entailment claims are conditional; their antecedents can be contingent like the instances of GS are.

with the A-facts. It offers a particularly clear conception of a reductive explanation. But it is not generally accepted.[15] I allow myself not to enter the debate since our topic is not whether and how meaning can be "naturalized." While, as far as I can see, all theories discussed in this thesis are compatible with it, we shouldn't commit conventionalists to Jackson's proposal. The deal I propose is this: If a conventionalist rejects Jackson's proposal, then she owes us a proposal to turn GS into a proper dependence-thesis; pending a word of protest, conceptual entailment is endorsed.

1.1.3 Normativity and semantic normativity

"Normativity" has rather opaque meanings and seems to lack a use in ordinary language. In theoretical contexts, something is called "normative" if there is an *ought* involved. To make the normativity thesis clearer, we need a better understanding of such *oughts*. I proceed by first considering a general notion of normativity and then turning to semantic normativity.

People generally agree that moral duties and prudential *oughts* are paradigm cases of *oughts*:

(5) **Moral duties**: We have a cleaning plan and I promised to do my part. According to the plan, I have to clean the kitchen today. So, I must (morally) clean the kitchen today.

(6) **Prudential oughts**: It's four o'clock and I am sitting in my office. I want to catch the 16:34 train to Duisburg. To do so, I have to leave my office now. Thus, I ought (instrumentally) to leave my office now.

Stemmer (2008:15) provides a particularly clear characterization of normativity in terms of what he calls "normative musts" (I typically use "ought" instead of "must"). I want to use a simplified version. According to his proposal, normative *musts* have the following three characteristics (my literal translation of Stemmer's proposal):

N1. It's a practical *must*. The object of the *must* are actions. Sometimes, the object can also be a state, the possession of a property, or the attaining of something, on condition that it is one's own acting that gets oneself in such a state or in the possession of the properties.

[15] Alternative positions are prominently discussed in the debate about physicalism, among them being *a posteriori* and non-reductive physicalism; see (Stoljar 2009:§8, §9).

N2. A normative *must* does not rule out that one acts differently than how one must. A normative *must* is not a force that inevitably moves (or will or would move) a person all the way to action.[16]

N3. Normative *musts* are tied to a pressure to act. It presses its addressee to do certain actions.

A few remarks about these characteristics are in order:

(i) Stemmer has in addition to N1–N3 also N4 which I *don't* want to endorse for the pre-theoretic characterization. For it seems that its sole function is to rule out certain (arguably implausible) conceptions of normativity according to which there are normative *musts* (or *oughts*) as ontologically basic entities *sui generis*:

N4. The normative *must* is always ontologically subjective. Its existence depends on thinking, feeling, and wanting of humans (or other living creatures).

(ii) What Stemmer considers to be *actions* in N1 is clearly more inclusive than actions. A better description for the object of a must is "something that a person can influence."

(iii) The pressure mentioned in N3 expresses a variant of motivational internalism according to which the following holds: If an agent believes that a norm to do X applies or accepts a norm to do X, then she feels a pressure to do X or is motivated to do X.

With these remarks in mind, I want to return to examples (5) and (6). Both are cases of a normative *must.* Yet, their *oughts* are of different kinds, as we know from the distinction in ethics between so-called "prudential" and "moral" *oughts*:[17] Prudential *oughts* have a *recommending* character while the *oughts* of moral have a *demanding* character. This distinction will be of some importance for my argumentation. Here I just want to observe that there is a clear distinction.

Semantic normativity Let us turn now to semantic normativity. I suggest to focus on selected communicative functions of meaning sentences.

[16] My wording is different from Stemmer. He writes that normative *musts* are not "action-determining". I prefer my wording which I borrow from Frankfurt (1971:8).

[17] See (Schulte 2009) for an elucidating norm expressivist analysis of the different kinds of *oughts* involved. The distinction is Kantian, *cf.* (Kant 1968; Pink 2004). I return to it in §8.2.1.

This helps us to find out whether we are somehow "talking *oughts*" when we are using these sentences. Some of the important communicative functions of meaning sentences are the following:

(7) To report an established (or conventional) use of an expression. Example: "zeitgleich" means *at the same point in time* among many people in Germany.

(8) To report a social norm about how one ought to use an expression. Example: "zeitgleich" means *in the same amount of time* while "gleichzeitig" means *at the same point in time* in German.

(9) To recommend a certain use of an expression. Example: When someone used "zeitgleich" but meant "gleichzeitig," one can utter "'zeitgleich' means *in the same amount of time* while 'gleichzeitig' means *at the same point in time* in German" to recommend a certain use.

(10) To demand a certain use of an expression. Example: When someone used "zeitgleich" but meant "gleichzeitig," one can utter "'zeitgleich' means *in the same amount of time* while 'gleichzeitig' means *at the same point in time* in German" to demand a certain use. The utterer of such a meaning sentence can only demand such a use if she has a certain authority over the addressee (*e.g.* being a teacher and the addressee being her pupil).

These uses of meaning sentences have different characteristics:

(i) If the first use is truthful, then there is a practice of using a certain expression in a certain way. The second use can be truthful without there being a practice of using an expression in a certain way. It's enough that there is a social norm according to which one ought to use it in a certain way. In contrast, there needn't be an *ought* involved for the first use.

(ii) In contrast with the latter two uses, the first two uses by themselves don't put the audience under pressure to behave in a way conforming to the practice or social norm, respectively. Such a pressure might still arise. *E.g.* if I want to behave in an inconspicuous way, then I react to a report about a practice or a social norm by adapting a conforming behavior. But the pressure wouldn't be present, if I hadn't also the desire to behave opportunistically. In contrast, the latter two uses create this pressure to conform whether I have such a desire or not (at least in normal circumstances where the hearer takes into account what the speaker has said and thinks that the speaker is sincere *etc.*).

(iii) While the first two uses can vary quite freely in the coordinate, the latter two cannot. I cannot recommend or demand how to use an expression in the past. This indicates that there is a difference between the first two uses and the last two uses.

(iv) The last two uses differ in their normative character – to recommend is something different than to demand. This seems to be the same distinction we observed above. (But on further inspection it is not. One doesn't make a moral demand if one demands to use an expression according to its meaning; see chapter 2 and §8.1.2.)

Let us focus on the fourth point. The fact that we can express *oughts* (or normative *musts*) with a demanding character by uttering a meaning sentence is central for the preliminary version of the normativity thesis N. Having recommendable character, I'll argue in chapter 2, is not sufficient. For what is recommendable to someone depends on her beliefs and desires, but I can be demanded to use an expression in a certain way, whether I want or not. Hence N is stated as follows as a conceptual claim:

N. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

If N weren't understood as a conceptual claim, it would be too weak for an authority relation can hold contingently. Hence, the *oughts* with a demanding character wouldn't necessarily depend on the fact than an expression means something. N deliberately leaves open in virtue of what there is an *ought*. In chapter 2, this question will be settled.

1.1.4 Research paradigms

Conventionalism about meaning has historically and systematically played an important role in philosophy and linguistics. Probably the first serious discussion of the thesis that meaning is conventional is by Plato in his dialog *Cratylus* dated from 360 B.C.E., where Hermogenes, one of the figures in the dialog, endorses a version of the conventionality thesis:

> I [. . .] cannot convince myself that there is any principle of correctness in names other than convention and agreement; [. . .] there is no name given to anything by nature; all is convention and habit of the users; – such is my view. (Plato 1969:384c–d)

Important for the understanding of Hermogenes' claim is that in those times, a convention was understood as a *verbal* agreement. Consequently we already need a meaningful language to express the agreement. But if the meanings in that language are also conventional, then a regress starts. This is a serious objection and will be addressed when I discuss Signaling Games in §5.3.3.

The early modern discussion improved the outlook. In his book *An essay concerning human understanding* John Locke introduced the notion of a tacit (or implicit) agreement which brings about a state of affair which is as if an explicit agreement has taken place.[18] But as Rescorla (2007) observes, the main difficulty is then to explain what it is to tacitly agree *as if.* Hume proposed in his *A treatise of human nature* that it is a system of mutual expectations and a common interest which induces people to behave in a way *as if* they agreed and which meets their interest.[19]

However, it was not before the twentieth century that the conventionalist project gained momentum. For until then, a better understanding of literal meaning was still missing. In particular Paul Grice's analysis of so-called "speaker-meaning" is notable. Speaker-meaning is a psychological notion defined in terms of beliefs, desires, and intentions. Grice thought that this notion can serve as a basis for theories about other kinds of meanings. He sketched in a series of papers an ambitious project: he wanted to analyze (literal) linguistic meaning in terms of regularities in speaker-meaning.[20]

Many current conventionalist proposals are strongly influenced by Grice's work. And so is the standard account which was developed by David Lewis in his book *Convention* (Lewis 2002). A popular version runs as follows: There are conventional regularities in the use and understanding of expressions among speakers and hearers of a linguistic community. These regularities determine what the expressions mean.

Today, there are three broad research paradigms:

Signaling Games (chapter 5) Signaling games form a family of formal models in game theory. They have been applied to the study of communication. The standard model was developed by David Lewis. It is arguably

[18] Locke expressed his view about tacit agreements in his *Second Treatise* of government, as a thesis by what money has value, namely by tacit agreement of men (Locke 1970:V.36–37).

[19] Hume's characterization of a convention can be found in (Hume 2003:III.II.II).

[20] A collection of Grice's papers on philosophy of language can be found in (Grice 1989a). He stated his ambitious project in (Grice 1989b).

the simplest model that is still interesting. According to it, speakers can send signals to hearers, who in turn can react to the signals. Certain conventional regularities amount to what we could call "successful communication." The model allows us to derive the signal meanings from such regularities. Usually, such an account is not proposed as a conventionalist account of meaning since it is too limited to explain the rich phenomena of our actual language use. For one thing, signals have no linguistic structure.

Actual Language Relations (chapter 6) Accounts of this kind also go back to Lewis. They employ a relation "Population P actually uses language $\mathcal{L}$" to assign a language (understood as a mapping from signs to meanings) to a community of language users. The relation holds iff among members of P certain conventions prevail. The conventional use typically consists in meaning and understanding expressions in certain ways.

Evolutionary Theories (chapters 7 and 9) In contrast to the accounts of the other two paradigms, Evolutionary Theories don't deem language use to be an Intellectual activity ("Intellectual" with a capital "I" is used as a technical term to characterize something as requiring rationality or being rational in a demanding way). The main proponent is Ruth Millikan. According to her, linguistic communication is not analyzed in Gricean terms.

1.2 Conventions and social norms

The word "convention" has many dictionary meanings – as in a "convention of people" (synonym: *meeting*) and the "National Convention" in France (the official name for the French government during the French Revolution between 1792 and 1795). The relevant meaning in our context is a different one which relates to a phenomenon which people call "customs," "traditions," or "established practices." These different words indicate that the phenomenon is not fully homogeneous. For the conventionalist project, a distinction between *conventions* and *social norms* is sufficient.

Conventions According to the sense of "convention" I want to explicate, the following famous cases from Lewis (2002:5–8) count as paradigmatic examples of conventions:

(11) You and I are talking on the phone and the connection is interrupted in the middle. We have a common interest to reestablish the connection. One of us has to redial and the other has to wait. If we both dial or both don't dial, then the interest is not satisfied. We both settle on on the "caller-redials" strategy.

(12) You and I sit in a boat and want it to glide in one direction while speed is not important to us. As long as we both row with the appropriate frequency and strength, we will satisfy our interest. There are many combinations of individual frequencies and strengths which are in this sense equally good. We both row in a way satisfying the shared goal.

(13) Car drivers share an interest in efficient and safe conduct on the streets they drive. Among the many ways they could behave when they drive, two of them are particularly simple: Crossing drivers drive on their respective right (or left – the second option). By doing so, they further their common interest. They drive on their respective right.

In these examples, the agents somehow manage to act in a way conforming to the convention. But their description leaves open *how* the agents act convention-conformingly, *e.g.* by deliberating about what to do or by habit. This lends itself to a natural classification of the class of conventions into three types of conventions, depending on the mechanisms sustaining them: (i) rationalistic conventions, (ii) rationally justifiable conventions, and (iii) dispositional conventions. Their names are suggestive: In case of (i), the agents rationally deliberate about what to do and this results in their convention-conforming behavior. In case (ii), the agents could deliberate but don't; rather non-deliberative mechanisms like habits elicit their convention-conforming behavior. Finally, the case of (iii) allows for non-rational agents that behave conformingly. This distinction will help us to evaluate different theories about conventions. Lewis' theory (chapter 4), for example, oscillates between (i) and (ii). Millikan's theory (chapter 7), in contrast, allows also for conventions of type (iii).

The examples have several important properties in common. They involve a group. The members of the group are interacting in specific ways with each other to satisfy a common interest: There is a pattern of activity which is realized and according to which some members behave. The structure of the situation is such that there is an alternative way to satisfy the common interest. The satisfaction of the common interest requires coordination. Whatever one does, it is only optimal to do so if other members of the group also behave in a suitable way. At least some uncoordinated ways of behaving lead to outcomes which are less optimal. This is an important trait of all examples: That what is optimal for a member to do depends on what the others do. Also, in the examples there was a regularity in the individual behavior. The regularities depend on the others acting suitably over some period of time. Acting according to the pattern has a self-enforcing effect. It is optimal for each agent who is doing her part to realize the full pattern to do so and this reinforces everyone to behave accordingly in new situations of a similar kind.

This description entails that the parties to a convention face a coordination problem. They "solve" it by having suitable behavioral dispositions[21] together forming a group-level "coordinative disposition":

A group has a *coordinative disposition* iff

1. its members behave in a way which tends to bring about a certain outcome which is optimal for them;
2. each of its members would behave differently if enough other members behaved differently;
3. if the members behaved as in description 2, they would behave in a way which tended to bring about a certain other optimal outcome; and
4. if a member behaves differently but the rest as in description 1, then the outcome tends to be worse for the deviating member.

The group level disposition is realized by suitable behavioral dispositions of its members. The individuals' behavioral dispositions share some of the marks of a coordinative disposition. In particular, (i) the members tend to have different behavioral dispositions iff enough other members had also different behavioral dispositions (in case of deliberation) *or* (ii) the members tend to have acquired different behavioral dispositions iff enough other members had also acquired different behavioral dispositions (in case of dispositions that are by and large innate).

Coordinative dispositions can be realized in different ways. "Dumb" agents could be guided by reactive behavioral mechanisms (think of the behavior of ants, *e.g.* the way they signal by using pheromones). In case of humans, arguably different mechanisms are at work. Their dispositions can be realized by sophisticated deliberation systems. They can also be realized by habits. Hence, optimal outcomes can be realized in different

[21] "Behavioral disposition" should be understood throughout the thesis in the wide sense including also dispositions to behave in a certain way – like coming to believe something – which, strictly speaking, are not dispositions to *act*.

ways. In case of non-rational agents, "being optimal" amounts to *being an evolutionary beneficial way of behaving in the situation the agent is in.* In case of rational agents, it amounts to *being a rational thing to do in the situation the agent is in.*

Let's call a coordinative disposition "effective" if it brings about an optimal outcome often enough. Such effective coordinative dispositions capture the core of conventions. In terms of them and the important properties pointed out above, we can characterize conventions as follows:

C0. A convention is *social*: there is a group G of agents.

C1. A convention *involves a pattern of activity*: (i) there is a pattern R of individual activities of members of G, (ii) R determines for a range of activities whether they are conforming or deviating, and (iii) at least on some occasions, members of G would behave in a way conforming to R.

C2. A convention *requires coordination*: Members of G have (together) an effective coordinative disposition to behave in a way conforming to R.

C3. A convention is *relatively robust*: in all near futures, it exists as well.

From conventions to social norms Not all cases that are ordinarily called *convention* satisfy neatly the characterization offered above. Consider again the rule of the road of many countries: to drive right on public roads. The case has some of the characteristic marks both of conventions and social norms. The case relates to conventions since the underlying strategic situation is a coordination problem which is characteristic for conventions. The case relates to social norms since drivers resent non-conforming behavior and there is a sanction mechanism – policing –, both being characteristic traits of social norms. This indicates that social norms should be distinguished from conventions, even if they are closely related to each other.[22]

In the literature, there are three conceptions: (i) conventions are social norms, (ii) social norms are conventions, and (iii) conventions are identical to social norms.[23]

[22] I'm not the first to observe this ambiguity of "convention". See *e.g.* (Kemmerling 1997:81–83) or (Hartogh 2002:viii) – but the way I explicate the difference is novel.

[23] Lewis (2002:97) endorses the first claim but surprisingly he also claims that conventions "may *be* a species of norms". Marmor (2009) and Glüer and Wikforss (2009) clearly endorse the first claim. Conventionalists about ethics endorse the second. Burke and

My proposal is to endorse a conception according to which conventions are not identical to social norms. There are at least three reasons for it: (i) Examples of what we ordinarily call a “convention” or “social norm” have different properties. I’ll turn to them below. (ii) Accepting my proposal clarifies the normativity debates about conventions (next paragraph) and meaning (chapter 2). (iii) My conception offers a theoretically fruitful perspective for the *use-determines-meaning* slogan. On the one hand, it allows us to clearly distinguish two cases: meaning in virtue of conventions and meaning in virtue of social norms (§1.3). On the other hand, it allows us to give a detailed answer to Humpty Dumpty’s problem since a social norm effects a social power structure among the parties to it while parties to a convention are all on an equal footing. (I return to the reason relating to the normativity debate about “conventions” in §8.2.3. I return to Humpty Dumpty in §9.4.)

The normativity debate about conventions All positions are taken in the debate about the normativity of conventions. Lewis (2002:97) holds that conventions are not normative. He faces the opposition of Margaret Gilbert (1989:§VI.5) and people following Hart’s analysis of a rule (Hart 1997). They claim that conventions *are* normative.[24] The position of Davis (2003:§9) amounts to the thesis that conventions are normative OR not (the definition is made flexible by using an OR-clause).[25]

It seems that these positions cannot be true together. But I think no one is simply wrong. I’d like to suggest that it is a verbal dispute. The reason is that they are not focusing on the same examples. Lewis, for example, is focusing on coordination problems where the parties’ interests in them seem to be exhausted by bringing about coordination. Gilbert, in contrast, is discussing various social settings in which there is a custom or practice where people have normative expectations directed to each other about what to do (and what not); her examples do not seem to be such that the parties’ interests are exhausted by bringing about coordination. But if so, the proponents in the normativity debate about conventions could all be correct if we distinguish the different cases, namely conventions, social norms, and normative conventions. I think that we should do so. (I return

Young (2009) endorse the third claim and thus also the first two.

[24] Among the “Hartians” are Glock (2010), Kemmerling (1976), and Savigny (1988).

[25] Davis’ position is not entirely clear. I think he is inadvertently noncommittal.

to this dispute in §8.2.3.)

I should say more about the relevant sense of *being normative* for a convention or a social norm. It's uncontested that conventions imply prudential *oughts*. Hence, we couldn't explain the dispute by understanding "normative" in this way. It's also not about moral *oughts*. For only a few conventions are morally relevant; this is already enough to rule out moral *oughts* as a candidate. I propose that a convention (or social norm) is said to be *normative* iff some party to it is in a position to demand conformity. It turns on the demanding-character some normative utterances about social norms have (see below). The proposal leaves open who is in a position to demand conformity; in particular it does not require that *every* party is in such a position.

Contrasting social norms with conventions The next four pairs of examples bring out differences between what we ordinarily call a "convention" or a "social norm." I first present the examples and then observe important differences between them. I suggest to call the a-examples "conventions" and the b-examples "social norms."

(14) Utility of conformity

a. Hume's boat: You and I sit in a boat. We want to get the boat move in one direction but the speed is not important to us.

b. The mischievous boss: In each staff meeting, the mischievous boss makes a nasty joke about his assistants. Since he has power, they are expected to play along nicely. But actually they detest his behavior.

In (14-a) conformity is individually beneficial on condition the other conforms as well. This contrasts with social norms. In (14-b), it does not individually make sense to conform.

(15) Existence of a good alternative

a. Battle of Sexes: A couple would rather spend the evenings together than do something separately. The options are going to football matches or going to operas. The wife would rather go to football matches than to operas; her husband has it the other way round. They convene on going to football matches.

b. Basic moral norms: Don't cause avoidable harm!

In (15-a), there is a good alternative, namely going to the opera together. A *good alternative* to some behavior is not just another pattern of activity; it's a pattern of activity that is in a sense equally good and could have prevailed as well. The existence of a good alternative is part and parcel of there being a convention. But in case of a social norm, a good alternative needn't exist. Examples are hard to find but basic moral norms, whatever they may be, are a case in point. Suppose, as in (15-b), not causing avoidable harm is such a basic moral norm. Then it seems there is no good alternative; for example, causing harm isn't one.

(16) Normative character and existence of sanctions

a. Signaling: People from Denver and Dallas want to recognize each other. To do so, the people from Denver wear a green button and those from Dallas a red button. Nobody demands that the others conform to the pattern. But it is individually rational to conform to the pattern.

b. Opening doors for women: People demand that men open the doors for women. Usually the men do so. Moreover, non-conformity is punished.

In (16-a), conformity can be recommended but one cannot demand that a party conforms (unless rationality is itself a norm). Sanction mechanisms need not exist for there to be a convention. In (16-b), conformity can be demanded. A sanction mechanism exists. Sanctions neither have to be severe nor formal – a certain look can suffice.

(17) Role of social structure

a. Signaling (again): People think about what is the thing to do and conclude that one ought to conform to the shirt-buttoning-convention.

b. Legislators and the police: The legislators thought hard about what the people ought to do and came up with a norm that determines what is allowed, prohibited, and the like, for a certain range of activities. The police enforce the norm.

In this last pair of examples, the role of social structure is at stake. In (17-a), social structure is not relevant. For there to be a convention, the existence of (a certain) social structure seems to make no difference. In (17-b), however, social structure is relevant. It has various roles, among them being: who is to act how, who is to sanction whom, and who is to decide how to act. In particular, we can distinguish three groups which are determined by such a social norm. First, there are the *enforcers* who enforce that the behavior of the addressees of the social norm conforms to it. Second, there are the *arbitrators* who are in a position to decide what it is to conform to (and to deviate from) the social norm. Third, there are the *addressees* who ought to conform to the social norm.

A pre-theoretic characterization of social norms The examples of what I call "social norms" are in some respects similar to and in others distinct from conventions. There are three similarities. (i) They are social in that they involve a group of agents. (ii) Social norms *involve a pattern of activity.* (iii) Social norms are also relatively robust.

But there are also differences between social norms and conventions. (i) A social norm does not need to have a point. That is, it need not individually make sense to conform to it. The case with the mischievous boss illustrates this. (ii) A social norm does not need a good alternative. The moral norm about avoiding harm is a case in point. (iii) Social norms have a feature conventions don't have: They prescribe, forbid, and/or allow certain certain conducts. (iv) In case of a social norm, conformity can be demanded, while in case of a convention, conformity is recommendable, it can not be demanded (unless rationality is itself a norm). This seems to be so because a sanction mechanism exists which enforces the accepted norm. (v) Social structure can be relevant for social norms, while not for conventions.

Informed by these considerations, I propose the following five properties to characterize social norms:

S0. A social norm is *social*: there are various not necessarily disjoint groups including (i) a group E of enforcers and (ii) a group G of addressees.

S1. A social norm *involves a pattern of activity*: (i) there is a pattern R of individual activities of the addressees, (ii) R determines for a range of activities whether they are conforming or deviating, and (iii) at least on some occasions, the addressees would behave in a way conforming to R.[26]

S2. A social norm is *prescriptive*: A social norm is, in part, constituted by a norm N which determines for a range of activities whether they are prescribed, forbidden, or allowed. The norm N prescribes to conform to

[26] By clause (iii) I rule out that there is a social norm that is never and would never be conformed to since I wouldn't want to call it a *social norm.* I think that it's a common feature of the examples I provided and for this reason we should make it part of the characterization. Otherwise, the point is purely terminological. So I won't argue for it.

R. *N* is enforced by the enforcers who accept it and have power over the addressees.

S3. A social norm has a *demanding character*: The enforcers are in a position to demand conformity to *R* from the addressees.

S4. A social norm is *relatively robust*: in all near futures, the enforcers accept the same norm and at least on some occasions, the addressees behave in a way conforming to *R*.

S2 might invite confusion. I distinguish between *social norm* and *norm*. They are used as technical terms (see chapter 8).

I offer a theory of social norms in chapter 8. I'll argue there that conventions and social norms can exist independently of each other. But their existence is compatible with each other. Moreover, I will also introduce an interesting hybrid notion: normative conventions.

Let's say (by means of stipulation) that there is a *stable social behavior* iff there is either a convention, a social norm, or a normative convention. A stable social behavior whose pattern of activity consists in using and understanding an expression is called a "stable use."

1.3 The case for meaning conventionalism

The viability of the conventionalist project depends on the conventionality of meaning. So, we better motivate the claim that meaning is conventional. My argument proceeds in four steps: First, by arguing that meaning globally supervenes on use; second, by arguing that this use is conventional; third by arguing that conventional use determines meaning; and fourth, by discussing meaning in virtue of social norms. The argument is not meant to be decisive but to motivate the conventionalist project.

First step: From use to supervenience Let us consider the following scenario:

(18) Suppose that for (almost) all English users the word "apple" meant *orange* and the word "orange" meant *apple*.

In (18), we seem to be inclined to say that these English users use the words "apple" and "orange" differently, namely "apple" when it's about *orange* and "orange" when it's about *apple*.

If so, then you accept the following (counterfactual) conditional:

(19) If among (almost) all English users the word "apple" meant *orange* and the word "orange" meant *apple*, then they would also use the words "apple" and "orange" differently, namely "apple" when it's about *orange* and "orange" when it's about *apple.*

But this is to say that a difference in meaning implies a difference in use. The claim can be generalized, since we have no reason to believe that anything about the scenario depends on the particular expressions or the language chosen. A reflection about what would be the case if we were to consider other kinds of words – adjectives, adverbs, verbs, ... – assures us of that. Hence, *meaning supervenes globally on use.*

Second step: From use to conventional use We have to show that the use in (19) satisfies the pre-theoretic characterization of a convention (§1.2): the use is social (C0), *involves a pattern of activity* (C1), requires coordination (C2), and is relatively robust (C3):

C0. The use is *social* since there is a community of language users.

C1. The use *involves a pattern of activity*: there is a specific pattern in the use and understanding of the words "apple" and "orange" and there are ways to use and understand them conformingly and deviatingly.

C2. The use *requires coordination*: The language users have an effective coordinative disposition. It's *effective* since they have a practice of using the words in a certain way. The description of the scenario is under an implicit *ceteris-paribus* clause: unless stated otherwise, things are as they actually are. Hence, the description of the scenario entails that the conditions of a coordinative disposition are satisfied:

   1. Using and understanding "apple" in a way that secures successful communication is optimal for us, not least because apples are relevant for us in our lives. They are liked and disliked, grow on trees around us and are healthy. It's beneficial to have a word for something that is relevant in our lives. So, we have reason to believe that having an expression to communicate about something which is relevant in our lives is generally something useful.
   2. If enough English users used and understood "apple" and "orange" as in (19), then others would follow their use.
   3. If we compare the use of "apple" and "orange" in (19) to their actual use in English, then it seems that both uses are equally optimal.
   4. If some deviated from the prevailing use and understanding of the

words, then successful communication would not be secured anymore. Hence, *ceteris paribus*, everyone would be worse off by deviating from the prevailing use.[27]

C3. The use is *relatively robust*: Given somewhat rational or stubborn language users, we have reason to believe that the prevailing use continues to exist in all near futures.

Therefore, the use of "apple" and "orange" is conventional. Again, nothing seems to depend on the particular words chosen. So, we can generalize to all expressions: their use is conventional. Since this is the use on which meaning globally supervenes, we can restrict the supervenience-base to conventional use facts. Hence, *the meanings of expressions supervene globally on their conventional uses.* Consequently, we have reason to believe that C is true:

C. For all expressions e, meanings m, coordinates $\mathcal{C}$: The conventional use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

There will be two changes to this thesis. A first one is to allow for meaning in virtue of social norms (next paragraph). The second is to address the meaning-without-use problem (think of very long or complicated sentences that would never be uttered but are meaningful nevertheless). The problem indicates that the supervenience-base consisting of facts about the conventional use of expressions is not large enough. For the time being, we may safely ignore this complication. I'll return to it in §6.1.5.

Third step: From global supervenience to conceptual entailment So far, a global supervenience claim has been established. But conventionalists want to make a stronger claim: Certain conventional uses of expressions in certain communities conceptually entail that the expressions have certain meanings among the members of the respective community. There are good reasons to believe that also the conceptual-entailment claim is true. For it is well supported: (i) By the above argumentation, a difference in the use of expressions implies a difference in their meanings. (ii) The particular conventionalist accounts that are discussed in the following

[27] The *ceteris-paribus* clause is necessary to make room for non-conforming uses; depending on the conception of these uses, occasional lies and non-literal uses count as such uses. They are tolerable if they don't occur too often and/or are detectable, *cf.* §6.2.2.

chapters explain how a proposition that an expression has a certain meaning is *a priori* derivable from propositions about its conventional use. (iii) The conceptual-entailment claim is further supported by my reflections on meaning in §7.3.2. (iv) The conceptual-entailment claim resists counterexamples and objections (see §2.1.2, chapter 3, and §6.1.5) – excluding the one to which I turn now.

Fourth step: meaning in virtue of social norms There is an objection against C: Some meanings of some expressions are not determined by conventions but by social norms. Since social norms are not conventions, this is incompatible with the truth of C. Hence, C has to be restricted. To justify these claims, let us consider some language-use scenarios:

(20) In many countries there are laws – implemented as social norms – concerning the correct use of the trade name "milk;"[28] violations are forbidden and can be punished. One can be demanded to use and understand the word accordingly. The addressees of the social norm are persons using the trade name "milk" for the purpose of doing business. The punishments (sanctions) create an incentive for them to conform to the regulations, that is, to apply "milk" only to milky liquids with at least 3% fat *etc.*

In this scenario, it seems that we're inclined to say that one of the meanings of "milk" is *milky liquid with at least* 3% *fat* among, say, *all* British people, even if the conventional use and understanding of "milk" in accordance with the meaning *milky liquid with at least* 3% *fat* is confined to a much smaller group of legislators who know the regulations and have power over all British people by means of a social norm. Plausibly, it's the use and understanding of these legislators (and not the use of all British people) which determines this meaning of "milk." A similar argumentation as the one offered above for meaning in virtue of conventions establishes this. Consider the following scenario:

(21) Suppose for the sake of the argument that (i) there is no convention concerning the use and understanding of "milk" and that (ii) its meaning were

[28] At the level of the European Union, the European Food Safety Authority makes recommendations for the member states of the EU relating to food products, see `http://www.efsa.europa.eu/`. At the national level, many countries have institutions with regulating power. *E.g.* in Germany, it is the Bundesamt für Verbraucherschutz und Lebensmittelsicherheit, see `http://www.bvl.bund.de/`. And finally, there is the law: the German law MilchFettG, §11(1), demands milk to have at least 3% fat.

milky liquid with at least 4% *fat* and non-conforming use effects demands to use it conformingly.

Plausibly, there would be a social norm regulating this use and understanding of "milk" which determines its meaning *milky liquid with at least* 4% *fat*.

The existence of a convention couldn't explain two marks of this case: First, "milk" has this meaning among *all* British people. Second, one can be *demanded* to use and understand the word in accordance with its meaning.

Two additional scenarios support the case of meaning in virtue of social norms:

(22) Many people in Germany use "zeitgleich" as a synonym for "gleichzeitig." However, "zeitgleich" means *in the same amount of time* while "gleichzeitig" means *at the same point in time* in German.[29]

(23) The word "camera" was once used in the sense *Apostolic camera* naming the (then existing) treasury department of the papacy. This is not common usage anymore but still listed as the first entry in Merriam Webster Online Dictionary (2010) and known to some, *e.g.* Hanks (2009:302).

In (22), the established (or conventional) use does not distinguish between the two meanings. Yet one can be demanded to use the words as indicated. Again, conventions cannot explain this.

With respect to (23), one who is using "camera" in the marginal sense of *Apostolic camera* does not seem to make any linguistic mistake. He can well be called a bragger or a weirdo for it. But this is neither to say that he isn't using a meaningful word nor that he's using it not in accordance with one of its meanings. It seems that "camera" still means *Apostolic camera*, even if there is no convention for it anymore. Moreover, it seems to be wrong to say: "Then it meant *Apostolic camera* but now it has lost this meaning." Rather, I think, we should say that in the 18th century "camera" didn't have that meaning and that nowadays, the word is ambiguous. In §7.3.2.2, I'll suggest that there are special social norms among us according to which it is allowed to use and understand an expression now as it was used and understood in the past.

[29] *Cf.* Sick (2004) who claims so. Some native speakers contest that the words have these meanings; this does not undermine the point since it is sufficient that it is a *possible* scenario.

So, the claim that there is meaning in virtue of social norms is justified. But then the conventionality thesis C should be restricted.[30] We can restate C as C′ in terms of stable use (including social norms and normative conventions which are a combination of conventions and social norms) as follows:

C. For all expressions e, meanings m, coordinates $\mathcal{C}$: The conventional use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

C′. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

C′ is supported by the argumentation above and addresses the difficulty resulting from meaning in virtue of social norms. Hence, we have a good reason to believe that it is true (with the *proviso* due to the meaning-without-use problem).

1.4 An adequacy condition

To evaluate conventionalist accounts we need an adequacy condition. The adequacy condition I propose consists of desiderata belonging to three groups: (i) Desiderata for an account of conventions, (ii) desiderata for an account of social norms, and (iii) desiderata for a conventionalist account of meaning. A conventionalist account is *adequate* iff the desiderata of (i-iii) are satisfied.

In short, developing a conventionalist account is a very ambitious project. It shouldn't surprise us if candidate accounts fail to be adequate. Satisfying some desiderata is better than satisfying none. When comparing different proposals, tricky issues could arise, if two candidates satisfied different but not all desiderata. Which should then count as better? I think that there is no answer in general. But this tricky situation does not obtain. The evaluation of the central proposals I consider is straightforward and yields a clear result.

Conventions An account of conventions has to satisfy the following desiderata in order to count as adequate:

[30] I discuss meaning in virtue of social norms in more detail in §9.3 on the basis of my account of social norms.

DesC1. The account must be faithful to the pre-theoretic characterization of a convention (C0–C3 in §1.2).

Of particular importance from the pre-theoretic characterization is that conventions are relatively robust. This is to say that it must be possible (i) that a system S of conventions with the goal g persists unchanged over time and (ii) that such a system can continue to prevail in a population in which non-conformance occurs quite often.

DesC2. The account must provide an answer to the question what conventions are (*e.g.* behavioral regularities, rules, patterns).

As we will see later, there is no agreement what kind of thing conventions are. Some say they are a kind of regularity in behavior. Others say that they are a sort of rule, and so on and so forth. Thus, for reasons of principle and clarity, it has to be said what conventions are.

DesC3. The account must provide a taxonomy of the kinds of conventions there are.

A taxonomy is part and parcel of good scientific practice. Of particular concern is the classification of linguistic conventions. In practice, I will restrict my attention to differences, if there are any, between conventions in general and linguistic conventions.

DesC4. The account must provide an answer to the question whether (and if so which) epistemic states are involved among the parties to a convention.

In the debates about conventions, one of the questions is whether conventions require common knowledge, mutual knowledge, and the like. It is contested whether parties to a convention need to be in certain epistemic state, and if so, in which kind of state exactly. Hence, an answer to this question is important.

DesC5. It must be possible that in a human population conventions are created, learned, sustained, and changed.

DesC5 makes an adequate account of conventions to be an account of conventions *for humans* – which is, after all, what we are interested in. While the topic of learning is very important, I won't discuss it in depth.

DesC6. The dynamics of conventions must be explained.

Conventions can be created (come about) which did not prevail before. They can cease to exist. Likewise, a system S of conventions with the goal g can change to a different system S' with the same goal g without breaking down during the transition. DesC6 requires an explanation of the basic operations of creation, death, and change of conventions.

Social norms The desiderata for conventions apply *mutatis mutandum* also to social norms, basically by substituting "social norm" for "convention." There is a complication: We can talk about "norms" in two ways, one of them being about norms *an individual* accepts, the other way being about *social* norms existing *in a group.* The thing to demand from an account of social norms seems to me that it accounts for both ways and puts them in relation. With this in mind, let me present the list of desiderata for social norms:

DesN1. The account must be faithful to the pre-theoretic characterizations of a social norm (S0–4 in §1.2) and of normativity (N1–3 in §1.1.3).

DesN2. The account must provide an answer to the question what (social) norms and normativity are.

DesN3. The account must provide a taxonomy of the kinds of (social) norms there are.

DesN4. The account must provide an answer to the question what kind of epistemic states are involved in a (social) norm.

DesN5. It must be possible that in a human population social norms are created, learned, sustained, and changed.

DesN6. The dynamics of (social) norms must be explained.

Conventionalist accounts of meaning A conventionalist account of meaning is adequate iff it satisfies the following desiderata:

DesM1. The account must be coherent with an explanation of the communicative functions of meaning-sentences.

DesM2. The account must explain the meanings of expressions in terms of their stable uses (conventions, social norms, and normative conventions).

DesM3. The account must provide a plausible notion of a semantic mistake.

DesM4. The account must allow for a plausible conception of a public language.

DesM5. The account must explain the usual meaning facts.

DesM6. The account must provide a solution to Humpty Dumpty's problem.

DesM1 and DesM2 are consequences of the task to provide a conventionalist account of meaning in the sense explained in §1.1.

Ad DesM1: Meaning sentences are, as we have seen in §1.1, of special interest.[31] The fact that they can be used to *recommend* and to *demand* a certain use of an expression is theoretically important. For they involve normative *musts* and have consequences for the explanatory architecture of the resulting account (*i.e.*, we need a notion of a social norm, see chapter 2).

Ad DesM3: We can make mistakes when we use a word not in accordance with its meaning. This fact is theoretically interesting. For making a mistake is something one ought not to do. An adequate account explains what it is to make such a mistake.

Ad DesM4: Public languages are up to a certain degree vaguely individuated; one word more or less does not make for a different language. An adequate conventionalist account must be coherent with such a conception of a language.

Ad DesM5: There are many interesting facts about meaning: some physical objects are meaningful; the semantic description of a language requires recursive semantic rules; there are systematic semantic relations between expressions; there can be expressions in languages which have a meaning but are not used; and more of the like.[32] A conventionalist account must explain such facts.

Ad DesM6: The way I'd like to understand Humpty Dumpty's problem is as follows. Humpty Dumpty claims that he can make words of a public language mean what he wants them to mean. *E.g.* "Apfel" actually means "apple" in German. But if Humpty Dumpty wants that "Apfel" in German means *orange*, then "Apfel" means *orange* in German. Alice is appalled and denies that he can do that. Both seem to have a point: Humpty Dumpty

[31] I ignore to a large degree questions about "higher-order" conventions depending on more basic conventions, like conventions of use as Morgan (1978) has introduced them, *e.g.* conventions about how to use words to perform so-called "indirect speech acts." I think that these questions should be answered in speech act theory, *cf.* (Staudacher 2007).

[32] *Cf.* (Lycan 2008:65).

in that it is we who make the words mean something; Alice in that a few members of a linguistic community are in general not sufficient to determine the meaning of an expression. Humpty Dumpty's problem is thus:

HD. If it's the use of expressions in a community which determines their meanings, then in which way does this determination depend on the members and the circumstances?

The problem is to provide an answer to the question. There are two guiding constraints on an answer. First, the determination seems not always to depend on everybody's use. Second, the determination actually depends on more than just a few (1, 2, 3, . . .) individuals. The first guiding constraint has the consequence that meaning in virtue of conventions cannot be the full answer to the problem. For then the determination would always depend on *every* party to the convention. (There is no notion of social structure used in the characterization of a convention.)

1.5 Overview

In this chapter, I've argued for a distinction between conventions and social norms. This has consequences for the conventionalist project. I've argued that expressions mean what they do in virtue of stable uses (conventions, social norms, and normative conventions). Hence, the explanatory architecture of an adequate conventionalist account must be so as to allow not only meaning in virtue of conventions but also meaning in virtue of social norms (and normative conventions).

In chapter 2 I discuss the normativity thesis. According to the proposal I defend, social norms – but not conventions – explain the normativity of meaning. Hence, whether an expression's meaning is normative or not depends on what determines an expression's meaning.

In chapter 3 I defend the conventionalist project against a fundamental objection from Davidson who argued that conventions are not essential for meaning. In the subsequent chapters, I critically discuss and improve on different conventionalist accounts.

In chapter 4, I introduce Lewis' theory of conventions and defend it against the known objections. Yet, the theory is not fully adequate since dispositional conventions cannot be explained by it. Lewis' theory is used in the two subsequent chapters 5 and 6 where I discuss conventionalist accounts of two paradigms: Signaling Games accounts and Actual Language Relations accounts.

Due to the lack of syntactic structure, Signaling Games accounts are too limited to yield an adequate account. Actual Language Relation accounts do better since they are sensitive to syntactic structure. But they have several inherent problems which are difficult to solve: (i) As standard Signaling Games accounts, they lack an account of social norms. Consequently, they can explain neither meaning in virtue of social norms nor semantic normativity, nor can they solve Humpty Dumpty's problem. (ii) Standard Actual Language Relation accounts are committed to the sentential primacy thesis. This is problematic. For one thing, it leads to highly underdetermined meanings for words. (iii) Standard Actual Language Relation accounts make linguistic communication too Intellectual. They define the use and understanding of expressions in terms of Grice's notion of speaker-meaning (and a Gricean notion of understanding). Hence, I plead for a theory of the third paradigm: Evolutionary Theories. In chapters 7 and 9, I turn to them.

In chapter 7, I present and evaluate the Evolutionary Theories of Millikan and Huttegger. They solve the third problem by not using Grice's notions; the other two problems reoccur.

In chapter 8, I offer an account of social norms. Thereby, the first problem remaining from chapter 7 is addressed.

In chapter 9 I offer an improved Evolutionary Theory. I follow Millikan to address the third problem. The second (and last remaining) is solved by treating sentences on a par with sub-sentential expressions. The "trick" is to define the meaning-determination claim directly for words and other kinds of expressions.

The thesis ends in chapter 10 with a positive conclusion: The conventionalist project makes a strong case for the conventionality of meaning. Different accounts remain viable; in particular my conventionalist account comes close to being adequate.

Chapter 2

Semantic normativity

> What *is* surprising is that despite extensive discussions of the topic it remains obscure exactly what the normativity thesis amounts to and why it should be endorsed.
>
> *Semantic normativity*
> Åsa Wikforss

"Semantic normativity," or the "normativity of meaning," seems to be one of these modern mantras in philosophy: repeated over and over but often devoid of any reportable content since the reader is not told what it is all about. Nevertheless, it's deemed to be a deep and serious insight about the nature of meaning. Some philosophers consider this feature of meaning even as a "litmus test"[1] for the adequacy of a theory of meaning. According to these philosophers, if a candidate theory of meaning does not pass the test, then it's no good in the first place. Case closed. That matters are not so easy is clear if we remind ourselves what it is to be normative or, at least, what it is for meaning to be normative. It's just not very obvious what an answer would be. To make progress, we first have to clarify what to make out of the talk about semantic normativity.[2]

[1] The expression is from (Wikforss 2001:203).

[2] Historically, the modern debates started probably with Wittgenstein's thesis that "the meaning of a word is constituted by the rules for its employment" (Wittgenstein and Waismann 2003:143). Subsequently, Winch (1963:33) pointed out that "the notion of following a rule is logically inseparable from the notion of making a mistake. If it is possible to say of someone that he is following a rule that means that one can ask whether he is doing what he does correctly or not." One way to understand this is that if there is a rule, then one can make mistakes by not following it; this is to behave incorrectly; one behaves incorrectly iff one ought not to behave so-and-so; hence there are the normative

On the basis of the characterization of normative *musts* in §1.1.3, we can say as a first approximation that semantic *oughts* are a kind of practical *must* which come with a motivational force to do what one ought to do. As such, they apply to action-like things: behaviors and reasonings. Moreover, as I observed in the introduction, the relevant expressions of the *oughts* have a demanding character: By expressing an *ought*, one can *demand* that something (not) be done.

But so far there is nothing *semantic* about the *oughts*. The obvious suggestion is to say that the *oughts* are about the meanings of expressions and that is what is semantic about it. But understood in this way, how could meaning be normative? Meaning is taken to be an abstract object defined by the roles it plays in theories of linguistic communication (§1.1.1). What could be action-like about meaning and involve an *ought*? This is the first question of this chapter and the answer I'll submit is that under certain conditions, meaning sentences can be used to express oughts with a demanding character. In §1.1.3, I put it in a tentative and noncommittal way as follows:

N. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

The demanding character of the expression of the *ought* should remind us of the distinction between conventions and social norms. While conventions necessarily have a recommending character, social norms need not (although some can have a recommending character). And while social norms necessarily have a demanding character, conventions do not necessarily have a demanding character (and if they have it, then it is in virtue of something else; see §1.2 and §2.2.3). Moreover, there is meaning in virtue of both conventions and social norms (§1.3). From this perspective, there seems to be an interesting connection between semantic *oughts* and the demanding character of social norms (and normative conventions). It suggests a the following conceptual claim (to which I return in §2.1.1):[3]

oughts. Kripke returned to the topic of rule following in his famous book *Wittgenstein on rules and private language* (Kripke 2000) where he endorses the slogan that meaning is normative (see p. 37).

[3] This also relates to a distinction made by Glüer and Wikforss (2009) between two kinds of theses about normativity which they call "meaning engendered normativity" (ME) and "meaning determining normativity" (MD). The normativity thesis N is about ME:

MD-ME. If meaning is normative, then this is because of the way it is constituted (determined).

Meanings of expressions are determined by conventions and social norms (in the sense of "determination" explained in §1.1.2, amounting to at least global supervenience, or as I prefer, conceptual entailment). Conventions have a recommending character. Hence, their *oughts* are not of the right sort for N. But the *oughts* of social norms are. So, plausibly, if an expression e means m at $\mathcal{C}$ in virtue of a social norm, then the meaning sentence "e means m at $\mathcal{C}$" can be used to express an ought with a demanding character. In contrast, if the expression's meaning is determined by convention, then its meaning sentence can only be used to express an ought with a recommending character. This proposal is further substantiated by providing an account of conventions (chapters 4 and 7) and an account of social norms (chapter 8 and §9.3).

A second question of this chapter is: Given an understanding of the normativity thesis, is the thesis true? As has been pointed out by Gibbard (1994) and Glüer and Wikforss (2009), the normativity thesis expresses a conceptual claim about the notion of meaning. Thus, N has a modal strength – it's claimed to be a conceptual truth (and not just a material truth) and N's truth in this sense is in dispute. I'll argue that N is too strong and hence false. But if N is appropriately restricted to cases of meaning in virtue of social norms, then it is true.

The third question is what the consequences are for the conventionalist project. I suggest answers to these questions on the basis of first considering arguments against and in favor of N (§2.1 and §2.2) – and then discussing the dialectic situation (§2.3). Thereby I execute the main argument of this chapter:

P1. There are good reasons to believe that meaning is not normative.

P2. There are good reasons to believe that meaning is normative.

normativity that results from meaning; in other words, ME is a thesis that relates semantic *oughts* to meaningful expressions and describes the character of these *oughts*. In contrast, the claim that social norms can determine an expression's meaning is about MD: (social) norms that are in some sense constitutive for an expression's meaning something. If the suggested claim is true (if meaning is normative, then it is because of the way it is determined), then the conventionalist project bears upon both ME and MD normativity: It explains ME in terms of MD.

C1. There are good reasons to believe that meaning is normative and that meaning is not normative.

P3. If an expression's meaning can be determined in different ways, one of them normative and another non-normative, then the tension in C1 is resolved.

P4. P3 is a good explanation for P1 and P2 and the tension that results from their conjunction.

C2. Meaning is determined or constituted in different ways.

P1 and P2 are established by the arguments against and in favor of N, respectively. C1 is a problematic intermediate conclusion which is simply the conjunction of P1 and P2. P3 is introduced as an explanation of what is going on. I consider some objections and suggest that P3 is a good explanation that is tenable. I think it's worth exploring it and will tentatively endorse C2. But if C2, then the conventionalist should develop a mixed C_+N_+ account which is both conventional and normative (§1.1).

Before I go on, I should adjust expectations. Anyone who has followed the debates about *semantic normativity* knows that they are difficult and that the claims are highly contested. In contrast with many other areas of modern philosophy, there is still no systematic treatment of the subject. For example, there are no textbook introductions or handbooks about it.[4] I won't close this gap. I suggest a way to think about semantic normativity with a very limited focus, namely on N. The proposal should be judged on its merits. There is also a pragmatic reason for my being opinionated and selective at places: the space of possibilities one should still organize and explore is too big. For this reason, I omit the question whether (semantic) normativity can be naturalized.

2.1 Arguments contra

Many accounts of meaning do not even mention normativity as a topic. Consider for example the following widespread linguistic frameworks: Lexical Functional Grammar, Head-Driven Phrase Structure Grammar, Categorial Grammar, or Chomsky's theories. In none of them does semantic normativity play a role which would be relevant for the normativity thesis N. And

[4] Only recently has there been an addition to the Stanford Encyclopedia of Philosophy, see (Glüer and Wikforss 2009).

the situation is similar in some areas of philosophy of language (*e.g.* the many semantic frameworks that deal with Frege's puzzle).[5] Is this attitude justified? Up to a certain point, it is. I consider two arguments against the slogan that meaning is normative. The first relates to an on-going debate around the views of Anandi Hattiangadi, the second to Akeel Bilgrami.

2.1.1 The Hattiangadi debate

In recent years, the slogan "meaning is normative" has been scrutinized by Hattiangadi, and later also Kathrin Glüer and Åsa Wikforss, opening a wider debate about candidates for the normativity thesis and their truth.[6] Their chief contribution was to ground the normativity debate initiated by Saul Kripke by formulating precise theses stating how meaning might be normative.[7] For Kripke (2000) never explained how his claim that meaning is normative is to be understood. Is it about speaker-meaning, meaning in an idiolect ("idiolectal meaning") or meaning in a public language ("linguistic meaning")? What kind of *ought* is at stake? What is its nature?

The contribution of Wikforss *et al.* should be understood in this context. I think that they object by and large rightly against the theses they consider. But it seems to me that they discuss the wrong kind of normativity theses. These have roughly the following form:[8]

(1) Necessarily, if S means m by e, then there is some action S ought (not) to do or some action S may (not) do.

(1) is not very precise about the kind of *oughts* involved. Several ways to make it more precise have been considered. Among them are the following two versions for its consequent:[9] (i) S ought to apply e to object o iff o is (has) feature f (yielding a thesis called "prescriptivity") and (ii) S applies e correctly to o iff o is has f ("correctness"). Hattiangadi rejects both

[5] I mean the puzzle that statements of the form "$a = b$" can be informative while "a" and "b" are coreferential.

[6] The first wave of articles consists in: (Hattiangadi 2006, 2007; Glüer and Wikforss 2008).

[7] They are, of course, not the only ones. Notable are also Gibbard (1994) and Millar (2002).

[8] More specific claims of this *form* can be found in, *e.g.*: Kripke (2000), Boghossian (1989), Boghossian (2005), Hattiangadi (2006), Glüer and Wikforss (2008). That the theses should have modal strength is, however, often not considered. An exception is Gibbard (1994).

[9] The names are from Hattiangadi (2006:224 ff.). A related discussion is also in (Hattiangadi 2007:51 ff., 181 ff.).

of these and so do Glüer and Wikforss. But rather than retracing their argumentation, I'd like to consider another question. Let us grant that they have successfully refuted (1) and its siblings. Does it follow that the normativity thesis N has been refuted, too?

I think that in an important sense the truth of (1) and its siblings are irrelevant for the truth of N. The reason is that theses like (1) are either about idiolectal meaning or about *(Gricean) speaker-meaning.* So, they are not about linguistic meaning as N is. Hence, even if the trio is correct, they reject the wrong thing. Thereby I do not want to say that the contribution of Glüer, Hattiangadi, and Wikforss is not important or commendable – I think the contrary for the reasons I've stated above. But still, they miss the target.[10] At least from the perspective of the conventionalist project, what counts is linguistic meaning.

This assessment might be too quick. I'll consider three moves on their behalf. The first move consists in saying that while (1) looks as if it were about speaker-meaning (or idiolectal meaning), it's really about the speakers' beliefs about linguistic meaning. But this move would miss the target, too. Suppose that meaning is normative. Then what a language user ought or ought not to do just doesn't seem to depend on what she believes. She would also violate an alleged semantic norm if she were misinformed about the meaning of an expression. Maybe it would excuse her behavior but it would be a mistake nevertheless. So, arguing against normativity theses which are about what speakers believe to be the case doesn't seem to be relevant for the question whether meaning is normative or not.

A second move can be put in question form: Wouldn't it be strange if speaker-meaning (or idiolectal meaning) were not normative but linguistic meaning were? After all these notions are semantic notions and thus belong to the same family. While I agree that this reasoning is *prima facie* plausible, I think that it is not very conclusive. With respect to speaker-meaning the reply is this: The impression that they belong to the same family is probably just a peculiarity of the English language. English just has the verb "(to) mean" to express both for a speaker to mean something and for an expression to be in the linguistic-meaning relation with a meaning. Other languages are not ambiguous in this way. German, for example, has

[10] Interestingly Whiting, who defends the normativity theses against the criticism of Glüer, Hattiangadi, and Wikforss, always states the antecedent in terms of *linguistic meaning*, see *e.g.* (Whiting 2007, 2009). This went unnoticed in the debate.

two verbs, “meinen” (for speaker-meaning) and “bedeuten” (for linguistic meaning), which cannot be used interchangeably. There are also obvious differences between the two notions. Speaker-meaning is analyzed as a relation in which a speaker and an utterance figures. In contrast, linguistic meaning features neither a speaker nor an utterance. Moreover, the relations obtain in virtue of different facts. For a speaker to mean something, having a certain complex intention while uttering something is sufficient. In case of linguistic meaning, this wouldn’t do and it’s even contested that such intentions are necessary (see §6.2.4 and chapter 7). The point is that in general, there is little reason to expect that if speaker-meaning has some feature, then linguistic meaning has it, too.

The case for idiolectal meaning is similar. A normativist is not committed to analyzing linguistic meaning in terms of idiolectal meaning. Moreover, an expression means something in a speaker’s idiolect in virtue of different facts than those in virtue of which an expression means something in a public language. For the former it’s a matter of the speaker’s psychology, while for the latter it’s a matter of the stable uses (convention, social norms, normative conventions) that exist in a community (§1.3).[11] Hence, the arguments by Wikforss *et al.* are at best incomplete. So, let us end in the same way we began the move, namely with a question: Why shouldn’t we expect further differences between speaker-meaning (or idiolectal meaning) and linguistic meaning, also with respect to their normativity?

Suppose for the moment that speaker-meaning is constitutive for linguistic meaning. If Glüer, Hattiangadi, and Wikforss are right, then speaker-meaning is not normative. This would be an important result. For, if linguistic meaning is normative, then its normativity cannot derive from statements about what a speaker means. This move is interesting. Consider, for example, a neo-Gricean account of linguistic meaning which analyzes linguistic meaning in terms of conventions about speaker-meaning. If Glüer, Hattiangadi, and Wikforss are right as we suppose, then the following is the case: If meaning were normative, then its normativity would have to be attributed to the conventions. For there seems nothing else in the account that could do the explanatory work. The reasoning behind this thought is the following claim (which I’ve already mentioned in the introduction to this chapter):

[11] I put here the issue of semantic externalism aside. It wouldn’t change the argument much and not a lot would be gained for considering the point.

MD-ME. If meaning is normative, then this is because of the way it is constituted (determined).

From a conventionalist perspective, this makes perfect sense. Linguistic meaning is not taken to be an unexplainable or unanalyzable notion. And it seems that if one claims that one can say in virtue of what an expression has a meaning, then one ought also to accept MD-ME. For then, it seems, all semantic features, including the normative, ought to be explained in terms of their determinants (*i.e.* stable uses of expressions). This brings me to my criticism of this move. Conventions are but one possible determinant of meaning we've considered so far, normative conventions and social norms being the other two (§1.3). Only the latter two have the relevant demanding character which can be used the explain the *oughts* of the normativity thesis. The question whether meaning is normative then depends on what determines a word's meaning. So, an argument would be needed to show why the determinants are necessarily of the relevant normative sort – or why they are not. (Bilgrami's argument below can be understood as providing this argument.) A second objection is that a Gricean speaker-meaning is not required to analyze linguistic meaning (as I've pointed out above). Consequently, for non-Gricean accounts, the arguments by Glüer, Hattiangadi, and Wikforss are not relevant. Thus, the third move fails to convince without further additions. (In fact, the alternative account I suggest in chapter 9 can be understood as supplying these additions without relying on Gricean speaker-meaning. The point *here* is that these additions are substantial and that Gricean speaker-meaning is not obviously constitutive for linguistic meaning.)

So, to conclude, while the Hattiangadi debate improved the dialectical situation by formulating precise normativity theses, the criticisms against these theses fail to be relevant for the normativity thesis N – with the exception of the last move turning on conventions and social norms, to which I'll return below.

2.1.2 Bilgrami's argument from regularities

Bilgrami (1993:129–138) argues against the normativity thesis by making the point that norms about meaning are not constitutive or in some other way essential for meaning because we can explain linguistic behavior without having to appeal to norms. Bilgrami does not deny that there are norms relating to meaning but he claims that we could do well without them. According to him, we need more than the existence of "extrinsic" (p. 129) norms to establish the normativity thesis.

In a first step Bilgrami points out that if meaning were to be normative, then there would need to be norms that have "[...] a high philosophical or constitutive profile, i.e. what might be called intrinsic norms" (Bilgrami 1993:129). Simplifying, we can say that the norms in question must be constitutive for the notion of meaning. In terms of an analysis of "meaning," the claim is thus:

P1. Meaning is only normative if one *must* analyze it in terms of norms about meaning.

But even if a norm pertains to meaning such as: "If you want to be understood, then you ought to use the words in their standard meanings," this doesn't entail that such a norm has a high profile (pp. 136–137). This raises the question how we can tell norms without such a profile apart from norms that have a high profile. Bilgrami proposes the following test for not having a high profile: If the candidate norm is derived from regularities in an individual's behavior, then it does not have a high profile. Or in Bilgrami's words:

> [I]f something which appears to be a norm is attributed merely on the basis of regularities in an individual's behaviour, then it is not a norm in any interesting sense. If it is not derived from or attributed on the basis of such regularities, if it is autonomous from such regularities, then it has more of a right to be called a norm.
>
> (Bilgrami 1993:129–130)

Bilgrami uses this test to argue that the meanings of lexical expressions are not constituted by such norms:

> But when it comes to the lexicon, it is perfectly possible to attribute concepts and meanings on the basis of observation of regularities in the behaviour of agents. As I have been saying there is no *compulsion*, therefore, to impose the norms dictated by social practice on an individual agent's concepts and meanings. We do have a choice in the matter. We do have a choice as to whether Bert has the concept of arthritis or tharthritis. We do have a choice as to whether KWBert has the concept of arthritis first and then later tharthritis or whether he has had the concept of quarthritis all along. It is possible to account

> for their behaviour no matter which of these alternatives we attribute to them. It is possible, therefore, to think of our attribution of such concepts as something done (holistically, of course) on the basis of observed regularities. Normativity, thus, simply does not have the same grip in this domain. (Bilgrami 1993:130)

I think one should distinguish between mental content and linguistic meaning. Bilgrami's examples about Bert and KWBert are about the attribution of *concepts*. But actually, he wants to make a claim about the *meaning* of lexical expressions. So, I think what Bilgrami wanted to claim is this: We can attribute meanings to expressions based only on observations of regularities in the linguistic behavior. In particular, for the attribution we don't require norms. These claims have their natural ontological counterparts: Regularities in the use of expressions determine their meanings, not requiring the existence of (social) norms. In terms of an analysis of "meaning," we can state the premise P2 of the argument and conclude:

P2. Meaning can be analyzed in terms of regularities in the behavior of agents without having to make use of norms about meaning.

C. Meaning is not normative.

The line of argument seems right if we put aside two complications, namely meaning without use (no use, no regularities!) and meaning in virtue of social norms. But we must strengthen P2. For regularities alone are not sufficient; they need to be *conventional* (at least in the weak dispositional sense; §1.2). First, it seems to be a conceptual truth about meaning that an expression could have meant something else.[12] But there just being a regularity in the use of an expression does not guarantee that it could have been different. For the regularity could be a law-like matter, like the thunder that follows the lightning. And we wouldn't want to say that thunder means *lightning*.

Second, if there is only a regularity in the use of an expression, it is possible that it ceases to exist in the near future. But it seems inconsistent to say that an expression means something but in its near future it means nothing anymore or something completely different.

Hence, we should strengthen P2 to P2′:

[12] The "could" is not to be understood in the sense that we could switch the meaning of a word at will – that might well not be feasible – but in the sense that things could have evolved differently.

P2′. Meaning can be analyzed in terms of conventional regularities in the behavior of agents without having to make use of norms about meaning.

P2′ is warranted by the argument for meaning in virtue of conventions (§1.3). Thereby, also the two objections can be addressed: First, the existence of a convention in the use of an expression entails that there is a regularity between the uses of the expression and some state of affairs.[13] Thereby, the expression gets its meaning (which is definable in terms of the regularity's state of affairs). Second, a convention has an alternative which could have prevailed equally well. Third, a convention continues to exist in its near future. Fourth and finally, conventional use patterns are patterns in which the participants communicate. Such conventions pick out behaviors which lead to successful communication. So, we have good reasons to claim that in virtue of such conventions expressions have a meaning.

This is still good news for Bilgrami. For conventions are not normative in the demanding-sense. Thus, together with Glüer, Hattiangadi, and Wikforss' arguments that speaker-meaning is not normative, there is nothing normative about conventional meaning. Hence, we have good reasons to believe that the normativity thesis N is false.

2.2 Arguments pro

In this section, I discuss three arguments that have been offered in support of the normativity thesis, namely: (i) Boghossian's argument from correctness, (ii) Itkonen's argument from common knowledge, and (iii) the argument from mistakes. The first is popular. The second is interesting since it relates to conventions. I argue that both arguments fail. The third finally establishes a weaker version of the normativity thesis.[14]

2.2.1 Boghossian's argument from correctness

There is a somewhat popular argument for normativity reasoning from an expression's having correctness conditions to the normativity of meaning. The classic statement is in Paul Boghossian's article *The rule-following considerations*:

[13] The nature of such regularities is the topic of §4.3.3.

[14] The list of pro-argument is not exhaustive; some additional arguments are mentioned in (Glüer and Wikforss 2009:§2.1.3).

> Suppose the expression "green" means *green*. It follows immediately that the expression "green" applies correctly only to *these* things (the green ones) and not to *those* (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely, that my use of it is correct in application to certain objects and not in application to others. [...]
>
> The normativity of meaning turns out to be [...] simply a new name for the familiar fact that, regardless of whether one thinks of meaning in truth-theoretic or assertion-theoretic terms, meaningful expressions possess conditions of *correct* use. (On the one construal, correctness consists in *true* use, on the other, in *warranted* use.)
>
> (Boghossian 1989:513)

The argument goes roughly as follows:[15]

P1. For any speaker S, expression e, meaning m, and object o: If e means m, then S applies e correctly to o iff o is (has) m.

P2. For any speaker S, expression e, and object o: S applies e correctly to o iff S ought to apply e to o.

C. For any speaker S, expression e, meaning m, and object o: If e means m, then S ought to apply e to o iff o is (has) m.

The argument is valid. P1 captures that meaningful expressions have correctness conditions. P2 expresses what it is to apply an expression correctly. C is the desired conclusion: in this sense, meaning is normative. Observe that C is ambiguous since the scope of "ought" is not fixed but the different precise versions don't change my evaluation.[16] It is generally agreed that P1 is uncontested. The contested premise is P2 in which correctness conditions play a central role.

The notion of a correctness condition for an expression is the one which we find in truth-theoretic semantics.[17] According to such a semantics, a

[15] Boghossian rejects the argument in (Boghossian 2003). Glock (2005:228–230) defends a similar argument. My objection against the argument is informed by Hattiangadi (2006, 2007) who raises similar objections. A systematic discussion of various versions of the argument can be found in (Glüer and Wikforss 2009:§2.1.1).

[16] The precise versions of C effect changes of P2 as well.

[17] In the debate, people discuss two conceptions of correctness conditions. I focus here on the conception of them as *truth conditions*. I do not discuss the construal of correctness conditions in terms of *assertability conditions*; again it's generally agreed that it wouldn't

predicative expression like "table" is applied correctly to some object *o* iff *o* is a table. So, the notion of correct application of an expression in P1 is the one of predication (Glüer and Wikforss 2009:§2.1.1). Likewise, a declarative sentence is applied correctly in a situation *s* iff the sentence is true in *s*. So, correctness conditions are nothing else but the familiar satisfaction conditions. The central job of these satisfaction conditions (ignoring complications like ambiguity, vagueness, and indexicality) is *classificatory*: They partition a space of things into two sets, say into the set of things to which the word "table" applies and into the set to which the word doesn't apply. One of these sets is labeled "correct application," the other "incorrect application."

In the quote, Boghossian seems to suggest that P2 is true since *correctness* is a normative concept. But if this justification is compelling at all, then it is only because it turns on an ambiguity of the word family "(in)correct(ly)". The word has at least two prominent uses, an evaluative and a classificatory one. Even if the evaluative one were sufficient to make the normativity thesis true, the one that is at stake is the classificatory one. But classifying something is not somehow to express an *ought*. It's just a separation of things. Something else is needed in addition: an *ought*, prescription, obligation, or the like. It's then this addition which yields an *ought*. In other words, the fact that an expression has correctness conditions by itself does not make the normativity thesis true.

What seems to be required in addition is something like: "One ought to apply words correctly!" And while this might be something which we generally demand, the demand is not obviously semantic anymore. One might generally demand that we ought to apply words correctly; for otherwise successful communication wouldn't be secured. But even if one granted such a proposal, the sources of the *oughts* are not of the right sort: they should be semantic but they are not. To summarize, so far it seems that the argument from correctness conditions to normativity is by itself incomplete and the obvious ways to fix the argument result in non-semantic *oughts*.

Recently, Whiting (2009) has defended the argument in ingenious ways. He observes that the anti-normativists grant a normative sense of "correctly" and then goes on in a sequence of steps to create room for the possibility that having conditions of correct application (P1) establishes that one ought

make a difference for the problems with the argument, see *e.g.* (Hattiangadi 2006; Glüer and Wikforss 2009; Whiting 2009).

to apply expressions correctly (P2). Therefore, he claims, meaning is normative (pp. 537–540). But as far as I can see, he's not making a compelling case for the positive claim that P1 (and hence P2) is normative in this sense. What he does is blocking the moves against the conclusion C and shifting the burden of proof to the anti-normativists (pp. 539 ff.). Whiting sees the debate as one between anti-normativists and normativists. He is simplifying matters. I, for one, am just interested in the argument and not interested in calling myself a normativist or not.

Whiting's "opponents" – Glüer and Wikforss in particular – disagree with Whiting's claim about where the burden of proof is. They think that the call is on the normativist (Glüer and Wikforss 2009:§2.1.1).

I think at this point we should simply concede that Boghossian's argument is not compelling for either side. It's too contested to settle matters. For this reason, I agree with Glüer and Wikforss (2009:§2.1.1) who write: "The debate over the simple argument might seem to suggest a basic clash of intuitions. The anti-normativist denies what the normativist asserts [. . .]."

2.2.2 Itkonen's argument from common knowledge

Esa Itkonen argues at different places that language is normative and that linguistics should be understood as a normative enterprise. From his recent book chapter "The central role of normativity in language and in linguistics" (Itkonen 2008), we can reconstruct the following argument:

P1. Common knowledge is constitutive for meaning.

P2. Common knowledge is normative.

C. Meaning is normative.

Let's grant that the argument is valid. I think we should reject it on the basis that there are good reasons to believe that the premises are not true. To begin with, the truth of P1 is contested. For example, if we endorse Millikan's account (chapter 7), then P1 is false. But at least within the Lewis/Schiffer/Bennett tradition, this premise is considered to be true. Thus, if one accepts the argument, then one commits oneself to a contested position about the role of common knowledge. But I am more interested in the premise which introduces normativity: P2. Why should it be true? Itkonen illustrates his point using a simple example about me (A) wanting to cash a check at a bank (p = that I want to ...) and a clerk (B) (p. 288). In such a case, Itkonen claims that A knows-1 p, A knows-2 that B knows-1 p, and A knows-3 that B knows-2 that A knows-1 p.

Itkonen now reasons as follows: "[T]hree-level knowledge of this kind necessarily occurs in all institutional encounters" (*ibid*; using language and meanings is such an encounter). Moreover:

> A's three-level knowledge about B is not about what B knows in fact, but what A is "entitled to expect" B to know: Given the surroundings, I was entitled to expect that the bank teller whom I was approaching knew his business, i.e. had the requisite three-level knowledge about me. Hence, common knowledge turns out to contain a *normative* element. (Itkonen 2008:288 ff.)

So Itkonen's claim is this: If there is higher order mutual knowledge of a proposition, then everyone in the group is entitled to expect that the others know it, too. But this still does not establish P2. The missing premise is that entitlements and (normative) expectations are clearly normative concepts and whatever necessarily involves such a concept is normative, too. This seems plausible.

Nevertheless, Itkonen's argument seems to be based on a misunderstanding. If we describe a situation to be such that there is some kind of mutual knowledge among members of a group, then we attribute a certain complex epistemic state to each member of this group. But from this attribution or the member's being in such a state it does not follow that they have certain rights or duties. Hence, Itkonen's claim that we are "entitled to expect" something is unwarranted. What he seems to mean is not common knowledge but rather *a system of mutual normative expectations*.

The misunderstanding is understandable. Lewis, who introduced the notion of common knowledge, used "expectation" in one of his definitions of a convention in a context relating to common knowledge.[18] But there are two senses of "(to) expect": (i) *(to) believe to some degree* (epistemic sense) and (ii) *(to) demand* (normative sense). The relevant one for mutual and common knowledge is the first one. Importantly, while one can expect something in both senses, one can expect something in the believe-sense without expecting it in the demand-sense. It seems that Itkonen did not pay attention to this distinction.

[18] See (Lewis 2002:78) and my §4.3.1.

Itkonen's argument would go through if institutional phenomena were constituted by such systems of mutual normative expectations. But is it really the case that necessarily, in an institutional situation, there is such a system? And is language in this sense still obviously an institutional phenomenon? I think not. The crucial problem is, however, that if one substitutes "a system of mutual normative expectation" for "common knowledge," then P1 presupposes in an uninteresting way what the argument should show.[19] Thus, I conclude that Itkonen's argument from common knowledge does not establish that meaning is normative.

2.2.3 The argument from mistakes

Another argument in support of the normativity thesis builds on the observation that we can make (semantic) mistakes when we use and understand meaningful expressions.[20] While it's not a knockdown argument, it's quite appealing. Consider the English users Sally and Ryan. Ryan wants some

[19] Suppose that (i) there is a social norm according to which one may demand that others have such common knowledge and that (ii) there is a social norm according to which one can be demanded to act rationally in making one's action depending on such common knowledge. If these social norms were also constitutive for meaning, then meaning would be normative in the required sense. But first, I think that these social norms are not constitutive for meaning because I hold common knowledge not to be constitutive of meaning in the first place (I argue against it in §4.3.1 and the accounts in chapters 7–9 show how meaning can be determined without assuming common knowledge). So, the assumption that there are such social norms and that they are constitutive for meaning would need to be justified and I don't think that such a justification is forthcoming. Second, meaning would then not be normative because of common knowledge, which is the argumentative idea Itkonen uses, but because of these additional social norms.

[20] The argument is based on an argument by Lars Dänzer developed in a term paper titled "How meaning might be normative." I had access to a draft version in February 2009. Many authors have discussed the connection between mistakes and semantic normativity, among them being Winch (1963:33), Kripke (2000), Wikforss (2001:209–212), Millar (2002:58–62), Glock (2005:228–230), Hattiangadi (2006:229), and Glüer and Wikforss (2009) – the final reference providing the most complete bibliography on the topic. A detailed evaluation of the different versions of these arguments is outside the scope of this thesis. Typically, the notion of a semantic mistake is based on correctness conditions and leads to a Boghossian-style argument (§2.2.1). Millar was one of the few who emphasized that there is another notion of correctness based on *using an expression in accordance with its meaning*. The argument offered here is based on his insight. Moreover, many arguments do not conditionalize on what an expression means in a public language to express what language users ought (not) to do with it. Rather, they conditionalize on what a speaker means/believes/intends (§2.1.1).

oranges and utters to Sally: “Sally, can you give me some apples?”. It seems natural to say: “Ryan made a linguistic mistake.” He used the word “apples” incorrectly for *oranges*. For “apples” means *apples* and not *oranges*. Because Ryan used the word incorrectly, he made a linguistic mistake. He used the word not in accordance with its meaning in English. I’ll call such linguistic mistakes that depend on an expression’s meaning “semantic mistakes.”

So, it seems that there is a close connection between (i) using a word in accordance with its meaning and using it correctly and (ii) using a word *not* in accordance with its meaning and using it *in*correctly. Counterfactual reasoning supports this claim: First, if Ryan had used the word “apples” for *apples*, then he would have used it correctly, and this is to say in accordance with its meaning. Second, suppose, as before, Ryan wants some *oranges* and utters to Sally: “Sally, can you give me some apples?” but Ryan didn’t make a semantic mistake. If this were so, then “apple” must have meant something else, plausibly *oranges*. We may summarize these considerations as follows:

P1. To make a semantic mistake is to use an expression not in accordance with its meaning (in a language).

Consider now a continuation of the initial scenario. Suppose that Sally gives Ryan what he asked for – apples – and Ryan reacts irritatedly by pointing to a bowl with oranges saying: “I wanted these!” Then, it seems natural and not inappropriate if Sally points out to him: “You said something else!”, or if she tells him: “You made a mistake because you said ‘apples’ ” or simply: “You are wrong!”, if she corrects him, or if she demands that he use the words correctly, namely by using the word “oranges” instead of “apples.” For the communicative purposes of Ryan, it would have been better if he used “oranges.” So, it would have been *optimal* (and *rational*) to do so. But it would also have been the *correct* thing to do. It was a semantic mistake. While they happen, they ought not to happen. One may expect that people use language correctly and demand such a use and understanding, at least in some circumstances. This suggests the following claim:

P2. One can be demanded not to make a semantic mistake.

Moreover, in the introduction (§1.1.3) I made the following observation about the communicative functions of meaning sentences:

P3. One can be demanded not to make a semantic mistake by someone's uttering a meaning sentence. Thereby, one expresses an ought with a demanding character.

The chain of reasoning is now as follows. By P1 to make a semantic mistake is to use an expression not in accordance with its meaning (in a language). Hence:

C1. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then one can be demanded to use e in accordance with its meaning m at $\mathcal{C}$.

C2. For all expressions e, meanings m and coordinates $\mathcal{C}$: If one can be demanded to use e in accordance with its meaning m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

C3. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

C2 follows from C1 and P3. C1 and C2 entail C3, which is identical to the normativity thesis N. While this argument seems plausible, it's too quick: it gives the impression that *anyone* can demand to use an expression in accordance with its meaning. But this seems wrong. Moreover, I didn't pay attention to the distinction between meaning in virtue of conventions and meaning in virtue of social norms. Hence, we will have to weaken the conclusion. Let us correct this by justifying the premises. I do so by providing an explanation.

Explanation of the premises According to the conventionalist project, uses of expressions that are conventional or governed by social norms determine their meanings. It seems that we should understand "use" here in the wide sense not only including a speaker's use of an expression but also including the ways hearers *understand* expressions. So, certain uses and understandings determine an expression's meaning. This suggests a way to explicate what it is to "use an expression in accordance with its meaning": A speaker S is using an expression e in accordance with its meaning in language $\mathcal{L}$ iff what S is expressing by using e is what e means in $\mathcal{L}$. Likewise, a hearer H is understanding an expression e in accordance with its meaning in language $\mathcal{L}$ iff, when H recognizes e as an expression of $\mathcal{L}$, H is understanding it in the meaning it has in $\mathcal{L}$. From this perspective, it seems plausible to say that one is making a semantic mistake iff one is using or understanding an expression of a language not in accordance with its meaning in the language.[21]

On this proposal, the truth of P1 and P2 is explained if the use of the expression in question is constituted by a social norm. For the existence of a social norm to use and understand an expression in a certain way implies that its enforcers can demand that an addressee of the social norm use and understand the expression conformingly. This explains who can demand of whom to use an expression accordingly. We can express such a demand by uttering: "You ought to use the expression so-and-so!" wherein the "ought" has a demanding character as required for the normativity thesis. Importantly, this is not implied if the use of the expression is constituted only by a convention. This is the reason why P1 has to be weakened, leading to according changes in the rest of the argument:

P1′. To make a semantic mistake is to use an expression – which has a meaning in a public language in virtue of a social norm – not in accordance with this meaning.

P2′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees not make a semantic mistake.

P3′. An enforcer can demand not to make a semantic mistake by uttering a meaning sentence. Thereby, she expresses an ought with a demanding character.

C1′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use e in accordance with its meaning m at $\mathcal{C}$.

C2′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If one can demand to use e in accordance with its meaning m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

[21] As Millar (2004:162 ff.) and Glüer and Wikforss (2009:§2.1.2) point out this covers both (i) performance errors (like slips of the tongue) and (ii) false beliefs about what an expression means.

C3′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use e in accordance with its meaning m at $\mathcal{C}$ (by uttering "e means m at $\mathcal{C}$" which expresses an *ought* with a demanding character).

C3′ is the normativity thesis that I endorse. But what's the status of P2 and P3 on this proposal? There are different notions of (semantic) mistakes and only on some of them are P2 and P3 true. We have to make sure that we use the same notion throughout the argument.

Semantic mistakes At this point I suggest that we take a step back and ask the question: "What is it to make a mistake?" By reflecting on our linguistic practice of calling things "mistakes," we gain an insight about what it is to make a mistake: When we call something a "mistake", then often it is because someone did something he shouldn't have done. So, one might give the following analysis:

(2) Agent A's behavior B is a *mistake* iff A should not do B.

(3) Agent A *makes a mistake* by doing B iff A's doing B is a mistake.

These analyses do not seem to be fully convincing. A better one would relativize them to *circumstances* and, maybe, take care of exempting factors. Let us assume that we can address these worries somehow and move on since nothing bears on these details. A more serious defect of (2) (and consequently (3)) is that it seems that there are behaviors B which A shouldn't do but, in some sense, A's doing B is still not a mistake. For example, there might be a moral (social) norm to marry before one engages in an intimate relationship and if so, it is a mistake to be in such a relationship while not being married. But yet, it might be the rational thing to do and not be morally demanded. In this sense, it can be that someone *should* not behave in a certain way while it's not a *mistake* to do so: The "should" is the moral *should* while the "mistake" relates to the *oughts* of rationality. The issue here is that we're changing the meanings of "mistake" and "should" in subtle ways. But they should be kept coordinated (indicated, say, by a shared index). Among the different ways to explicate the "should" in (2) are:

- Instrumental rationality: All things considered, A should do behavior C rather than B since C's expected utility for A is higher than B's expected utility for A.[22]

- *Ought* of a social norm: It's a social norm (in a relevant situation and for A) not to do C.

Before I argue for a notion of a semantic mistake in terms of *oughts* of social norms, some remarks are in order. First, we can distinguish different kinds of mistakes by distinguishing the different ways to understand the "should" in (2). In other words, the different "sources" of *oughts* correspond to different kinds of mistakes one can make.

Second, the list is incomplete; *e.g.* there are things one should (not) do for aesthetic reasons, there are things which violate the norms a person *individually* accepts, there are things that morality demands, and there are things which go against the teleological purposes of something. However, these don't seem to be relevant for the explication of the ought of a semantic mistake. Recall that the *oughts* in question (i) should be *semantic* in the required way and that they (ii) should have a demanding character. These conditions are not satisfied by these other kinds of *oughts*. For example, while it's wrong to make a semantic mistake, it's in general not wrong in the moral sense; so moral demands are not semantic in the required way. Neither are personal norms. For semantic mistakes depend on an expression meaning something in a *public language* (P1′). This is, obviously, a social affair while individually accepted norms are not necessarily so. This judgment is confirmed by the following conditional: Even if one used "entomologist" for *etymologist* in English and thought one ought to use and understand "entomologist" in this way in English, one would be making a semantic mistake in English. It seems natural to accept the conditional. By doing so, also personal norms can be ruled out since they are not semantic in the required way for meaning something in a public language.

Third, the *oughts* of conventions are prudential *oughts*, the *oughts* of instrumental rationality. They have a recommending character – but not necessarily a demanding character.

Fourth, it seems to me that prudential *oughts* are only *oughts* with a demanding character in virtue of a further *ought*: the *ought* to be rational. This *ought* is a very general one and at least in many situations we want to be rational and our peers want and demand us to be rational.[23] The

[22] Expected utilities are introduced in §4.2.1. The things taken into consideration can include what someone else wants A (not) to do.

[23] Schulte (2008:207–224) develops and defends this position in detail.

point is that just because an action has a maximal expected utility for an agent, there is no *ought* with a demanding character to do the action. Only if there is a further *ought* to be rational, is there such an *ought.* Arguably, there must be a social norm for this.

My suggestion is that the notion of a semantic mistake which is used in the argument is the one that is constituted by social norms. My argument proceeds by elimination of the other proposals and makes use of the observation that if someone makes a semantic mistake, then one may correct him and demand that he use the expressions correctly. That one may do so is not to say that one must do so. We correct each other and make such demands only under special circumstances, *e.g.* when we are in a language learning situation. Among so-called "competent" language users this kind of behavior occurs much less. But this is not decisive for the matter. What is important is whether one may rightfully behave so, even if one is thereby behaving in a nitpicking or bossy way. That the ones who made a mistake perceive the situation so when they are corrected or demanded to talk correctly might also be the explanation why adults don't correct that often: politeness demands not to.

This leaves the two kinds of mistakes listed above: mistakes of rationality (or mistakes of conventions), and mistakes of social norms. Again, this is not to deny that there may be other kinds I haven't considered but it seems implausible to me that they should be relevant for what is at stake. The two options are related to the ways an expression's meaning in a language is determined. Above, we've observed that whether a use of an expression is a semantic mistake or not depends on its meaning in a certain language. On a conventionalist account, meanings are determined by stable uses (convention, social norm, normative convention; §1.3). So, what one should say is that whether a use of an expression is a semantic mistake or not is determined by *that* which determines the meaning of the expression in question – essentially a convention or a social norm. Corresponding to these, there is a difference with respect to the normative character they have. Only social norms (and normative conventions which are in part constituted by a social norm, see chapter 8) imply that some members of the groups in which they prevail are in a position to demand conformity.

Moreover, if there were a linguistic convention, then a member would behave foolishly or stupidly or against her interests by not using an expression in a way conforming to the convention. But this is something else than to make a semantic mistake. It seems to me that the instrumentalist loses this distinction. This is another argument against analyzing the notion of a semantic mistake in terms of prudential *oughts*.

Hence, it follows that semantic mistakes are constituted by social norms. The reasoning is this. First, normative conventions are constituted by social norms and the demanding character they have is due to their social norm. Second, as observed in the first step, at least sometimes, one can demand from someone to use and understand an expression in accordance with its meaning. Therefore, by inference to the best explanation, semantic mistakes are constituted by social norms.

There is one caveat. In the second step, I've observed that it can be the case that one ought to be rational. If this is the norm, then one can be demanded to be rational. One could thus object to the conclusion just reached that the "semantic" demands to use and understand expressions in accordance with their meanings needn't be explained in terms of social norms but could also be explained in terms of conventions, or more generally, in terms of instrumental rationality. This objection can be answered, I think. First, when I considered that to be rational "is the norm," then "the norm" must be understood as a *social norm*. The entomologist/etymologist case should have convinced us that the relevant *oughts* cannot be personal. Thus, semantic demands are again explained in terms of social norms. For, we've just added an intermediate means-end step. One would say now that to use and understand expressions in accordance with their meanings is the rational thing to do and this is what one ought to do because there is a social norm to do so. There is a second rejoinder: It just doesn't seem to be *prima facie* plausible to claim that every use and understanding of an expression in accordance with its meaning is the rational thing to do. Maybe the claim can be defended against all the cases where one might, at first, think otherwise. But it seems to be an open question whether such a defense succeeds. For this reason, it seems to me, the explanation offered above which was directly in terms of social norms is in the better position to be true.

Summary The upshot is that if one can make semantic mistakes by using a certain expression, there is a (linguistic) social norm (relating to it). We can make semantic mistakes. Thus, there are (linguistic) social norms. In contraposition this means that if there are no social norms, one cannot make semantic mistakes. In particular, if there are only conventions, then one cannot make semantic mistakes. One can make a mistake of rationality but this is something else – it lacks the demanding character semantic mistakes have. Closely related to making mistakes is to correct someone for uttering something which was not in accordance with the meanings of the words used. *Mutatis mutandum*, correcting a semantic mistake also depends on there being certain (linguistic) social norms.

Thus, to conclude, insofar as an expression means something in virtue of a social norm, one can make semantic mistakes by using it. In this sense, the argument establishes that meaning is normative. But insofar as a meaning of an expression can be determined solely by a convention (and not also by a social norm), the argument does not apply. I take the plausibility of the conventionalist accounts in the subsequent chapters to be sufficient to establish that a meaning of an expression can be determined solely by a convention. Moreover, Bilgrami's argument established that conventional meaning is not normative (§2.1.2). For this reason, the conclusion should be: the normativity thesis N cannot be true in general since its truth depends on the expressions' having meanings in virtue of a social norm and that's not guaranteed because of the possibility of meaning in virtue of convention.

2.3 Discussion

The critical discussion of the arguments against and in favor of the normativity thesis raised issues to which I'd like to return now. First, we should establish the dialectic situation and thereby return to the main argument I presented in the introduction to this chapter. I use the opportunity to suggest an interpretation of what is going on and to return to the questions of this chapter (the normativity thesis? its truth? its consequences?).

P1. There are good reasons to believe that meaning is not normative.

P2. There are good reasons to believe that meaning is normative.

C1. There are good reasons to believe that meaning is normative and that meaning is not normative.

P3. If an expression's meaning can be determined in different ways, one of them normative and another non-normative, then the tension in C1 is resolved.

P4. P3 is a good explanation for P1 and P2 and the tension that results from their conjunction.

C2. Meaning is determined or constituted in different ways.

The summary of the argumentations in the previous sections is as follows:

Contra arguments: The Hattiangadi debate reminded us that we should focus on linguistic meaning. In the discussion, we arrived at the insight that if meaning is normative, then this is because of the way it is determined (MD-ME). Thereby a relation is established between (i) what determines an expression's meaning and (ii) its normativity. In short: *That* in virtue of which an expression means something is *that* in virtue of which utterances of meaning sentences can be used express an *ought* with a demanding or recommending character. Bilgrami's argument from regularities established that conventional meaning is not normative and hence P1.

Pro arguments: Boghossian's argument from correctness is not compelling. Neither was Itkonen's argument from common knowledge. But the argument from mistakes established that meaning in virtue of social norms is normative and hence P2.

The dialectic situation is now that we have good reasons in favor of and against meaning being normative (C1). But there is no contradiction. The reasons in favor of and against meaning being normative are conditional on an expression's meaning's determinant (conventions, social norms, normative conventions). This observation is premise P3 of the main argument. This goes against an implicit assumption that seems to be part of many foundational meaning theories. According to them, the task is to come up with an explanation why (in virtue of what) expressions mean what they do and the explanans is taken to be something *uniform* for all expressions, for example conventions, rules, intentions, mentally realized grammars, ... It seems to me that this assumption is not well supported.

There being such a uniform explanation would be appealing. But why should we think that there is such an explanation? The argument that meaning is conventional in §1.3 suggests that use theories of meaning are on the right track. From that perspective, it should not come as a surprise that the ways languages can be used are multifarious. The argument also supports the claim that the meanings of expressions should be explained in terms of conventions or, distinctly, in terms of social norms. Moreover, in §7.3.2 and §9.3.1 I'll discuss in depth several cases which strongly suggest that there is not a single "thing" that explains why expressions mean what they do. For these reasons, I think that the null-hypothesis should be that there is no uniform explanation of the "meaning in virtue of ..."-question. This is consistent with what P3 says. The justification of P3 is by means of its success: It explains the seemingly contradictory positions on semantic normativity.

Moreover, by distinguishing between conventions and social norms we can also explain our ambivalence in endorsing and rejecting the claim that meaning is necessarily normative.[24]

Let us consider an objection against P3: only (certain) normative conventions determine an expression's meaning. Consequently, meaning is normative. The impression that there can be meaning in virtue of conventions is wrong. If there were such non-normative conventions, then the expressions wouldn't mean anything. Moreover, we have the impression that it can be that there is meaning without normativity because semantic *oughts* can be overridden. The claim about meaning in virtue of non-normative conventions boils down to an intuition: Are we inclined to say that a linguistic expression can mean something and yet there are no semantic *oughts*? I'd say so and so would the adherents of the subsequent accounts in the next chapters. For they make a compelling case that non-normative conventions can determine an expression's meaning. Part of my reason to be so inclined is that the notion of meaning I'm interested in is defined by its roles in a theory of linguistic communication (see §1.1.1). It seems to me that such a notion needn't necessarily be related to *oughts* with a demanding character. This is not to claim that it can't or shouldn't be so related. Rather, it seems to me that we should allow for this possibility as well. But if one accepts my position that an expression's meaning can be determined or constituted in different ways, then the objection fails.

So, it seems that P3 is a good explanation (P4) and at least tenable. I don't want to claim that it's the best explanation. But since it offers an interesting perspective, I think we should tentatively endorse it and develop the consequences. This is what I will do. Hence, I accept the thesis that an expression's meaning is determined or constituted in different ways (C2).

[24] This is one of the reasons why we should separate conventions from social norms; §1.2.

2.4 Summary

Let us return to the three initial questions of this chapter. The first concerned the normativity thesis and what about meaning is action-like and involves an *ought*. The argument from mistakes suggested an answer: It's using an expression in accordance with its meaning that is action-like. In case of meaning in virtue of social norms, one can be demanded to use the expression accordingly. This is about meaning since (i) using an expression in accordance with its meaning consists in using and understanding the expression in certain specific ways and (ii) such use-patterns, if they are conventional or governed by a social norm, determine the expression's meaning. The key insight for (ii) was that if meaning is normative, then this is because of the way it is determined (MD-ME).

On the basis of the arguments considered, we have good reasons to claim that the meaning of an expression is not normative if it means what it does in virtue of (non-normative) conventions. Hence, we should abandon N and accept the revised N′:

N. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character.

N′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use e in accordance with its meaning m at $\mathcal{C}$ (by uttering "e means m at $\mathcal{C}$" which expresses an *ought* with a demanding character).

To make an even stronger case for N′, I provide an account of social norms in chapter 8 which is used to explain semantic normativity in §9.3. This brings me to the second question concerning the truth of the normativity thesis. I've argued that N is false and that N′ is true. For this reason, it seems that the *causa* of semantic normativity should not be considered to be a "litmus test." The third question was about the consequences for the conventionalist project. In short, what we need is an account according to which an expression can mean something in virtue of conventions or social norms and that entails N′ is true (*i.e.* a C_+N_+ account; see §1.1).

2.5 Using language solely in virtue of conventions

The distinction between conventions and social norms has further consequences: If only the former existed in a population using a language, then their use would be limited. To overcome these limits, social norms are required – or so I'll argue. While this topic is not directly related to *semantic* normativity, it highlights the explanatory limits of a conventionalist account that lacks a notion of a social norm.

Let me start by being more precise about what the difference is between *using a language solely in virtue of conventions* (for short: C-using a language) and *using a language in virtue of conventions and social norms* (for short: using a language). Once we have a better understanding about this, it becomes evident what plausible explanations could be. Consequently, I'll elaborate on what I mean by "C-using a language" and "using a language." This allows me to observe certain limitations of the former that the latter doesn't have.

In a first approximation, we can say that the C-use of a language is one that is not constituted by social norms. It's thus conceptually possible for beings not capable of having social norms to use language in this way. In other words, there is a conceptually possible world, call it w_c, in which members of a community use a language but there are no social norms among them. In contrast, the use of a language is one which is at least partly constituted by there being social norms. Consequently, the members of the community in w_c couldn't use a so constituted language.

Along these lines, "the use of a language is constituted by social norms" means at least this: There must be certain social norms for certain purposes. Whatever the purposes of language are, one is simply to use it. To use language is to perform speech acts, or in Austin's words "to do things with words" (Austin 1975). Consequently, the distinction we're exploring can be analyzed in terms of actions one can perform in virtue of there being social norms which one couldn't perform if they didn't exist.

With this in mind, let me present some ordinary things one uncontestedly can do by C-using a language:

- Speakers can use expressions in accordance with their conventional meanings as parts of basic speech acts of expressing attitudes like beliefs, desires, and intentions.
- Hearers can understand expressions in accordance with their conventional meanings as parts of the same speech acts.

- One can express one's judgment to speakers and hearers whether their behavior was rational or not. In particular, one can express one's judgment that it was not and one can express one's desire that they behave conformingly. (We might call such an act a "quasi-correction.")

In contrast, there are some ordinary things one can't do:

- To make a linguistic mistake
- To correct someone (in the strong sense of the word)
- To promise something to someone
- To inform someone about something

I've already argued for the first two claims in §2.2.3. The last two claims have still to be established.

Part of what it is to (sincerely) promise to do some action *a* is to intend to undertake the commitment to do *a* (Searle 1969). For someone to undertake a commitment is more than her having certain intentions to do something. For she can also be blamed for not living up to what she has herself committed to. Moreover the addressee can claim a right. So, plausibly, to undertake a commitment is to be analyzed in terms of a social norm that redistributes rights and duties by certain means. The social norm could be one whose pattern of activity is for the speaker as follows: if she produces an utterance of the form "Hereby I promise to do *a*" under certain circumstances, then she undertakes a commitment to do *a*. Enforcers of the social norm can demand conformity to this pattern. Typically, her hearers are among them. This is why the addressee of a promise can rightly protest if the speaker promised to do *a* but didn't do it and that's why the speaker should accept the blame.[25]

The case for *to inform someone* is more contested. Eike Savigny (1983, 1988) explicated what it is *to inform someone that something is the case* in terms of redistributing rights and duties between the speaker and the hearer.[26] The rights and duties are redistributed in virtue of there being social norms. In particular, the redistribution consists in the speaker permitting the hearer to rely, at the expense of the speaker, on the speaker's

[25] Observe that by claiming that we need social norms for promises one does not make a claim whether it is rational to keep a promise or not. I think it is but this is a different issue.

[26] In this respect, von Savigny's proposal is similar to (Brandom 1994).

claim that *p* (Savigny 1988:§5).[27] Thereby, a hearer that has been informed that *p* by a speaker, in case that not-*p*, is in the position to demand compensation from the speaker.

I think von Savigny's explication does not capture all kinds of informing someone. Yet, I believe that it's an ordinary sense of *informing someone.* An influential counterproposal is the theory developed by Kent Bach and Robert Harnish in their book *Linguistic communication and speech acts* (Bach and Harnish 1980). According to them, to perform a speech act is (just) to express a certain attitude, at least for a broad class of speech act types (but not all, *e.g.* not the legalistic ones considered above). For example, to *inform* the hearer that *p* is to express the belief that *p* and to express the intention that the hearer form the belief that *p* (pp. 39–42).

Given the simple conception of speech act types of Bach and Harnish, one might well wonder who's right about the case of informing someone: Is it von Savigny or Bach and Harnish, or yet someone else? As von Savigny points out in connection with his explication of asking, it's hard to argue that some particular analysis of a speech act is adequate or the right one (Savigny 1983:§35). In part, this is so because it's an empirical matter what individuals mean by, say, "informing" and whether all individuals share a meaning corresponding to one of the proposals, be it von Savigny's, Bach and Harnish's, or someone else's.

There seem to be ordinary cases where someone expresses what Bach and Harnish claim, namely a belief that *p* and the intention that the hearer form the belief that *p*, *but yet don't inform someone.* The point is that the expression of a belief and of a certain intention is not sufficient for performing the speech act of informing. For example, if you ask me whether the swimming pool at the other end of the city is open and I inform you that it is, then it seems part and parcel of my informing you that you may rely, at my expense, on the claim that the pool is open and that I owe you something, if you cycle to the other end of the city to find the swimming pool closed. I find such cases of *informing someone* important and natural uses of language. Moreover, on the one hand, the fact that I have certain

[27] Von Savigny uses the term "convention" where I use the term "social norm." Using my terminology, I think my use of "social norms" for von Savigny's "convention" is justified since what von Savigny calls "conventions" is a version of Hart's analysis of a rule (Hart 1997) which does not differ from my analysis of a social norm in the relevant respect that it has a *demanding character.* That is, that some members of the relevant group in which the social norm prevails can demand something in virtue of its existence.

duties by having informed you and the fact that you thereby have certain rights are not explained by Bach and Harnish's proposal. On the other hand, von Savigny's proposal explains these facts nicely. This suggests that informing someone in the sense of our explication is on the right track and should be considered to be ordinary.

The crucial point is the redistribution of rights and duties. While it can directly be explained in terms of social norms, the alternative account by Bach and Harnish cannot. For this reason, also an alternative explanation in terms of the speaker's credibility wouldn't work. While it can explain why a hearer should or shouldn't trust the speaker's information, it cannot explain in virtue of what the speaker has incurred a duty and in virtue of what the hearer has earned a right.

Moreover, it seems impossible to change Bach and Harnish's proposal in a way that both avoids social norms and yet explains the redistribution of rights and duties. The obvious move which stays true to their central idea that to perform a speech act is to express an attitude would be to make the expressed attitude more complex. One could propose that S's informing H that p also consists (i) in S's expressing her belief that H may rely, at her expense, on the fact that p is the case and (ii) in S's expressing her willingness to pay for H's expenses if p is not the case. Such an amended explication of informing someone seems to me to be implausible. First, it seems wrong to say that one expresses these attitudes when one informs someone. Second, while it is now the case that the hearer is paid for her expenses if p is not the case, it's still not explained that the hearer has such a right. For these reasons it seems not possible to get a Gricean proposal of informing someone with the features it should have.

Chapter 3

Davidson's malapropisms

> [W]e should try again to say how convention in any important sense is involved in language; or, as I think, we should give up the attempt to illuminate how we communicate by appeal to conventions.
>
> *A nice derangement of epitaphs*
> Donald Davidson

In this chapter, a standard objection by Davidson is the topic. Famously, Davidson argued against linguistic conventions on the basis of a phenomenon called "malapropisms".[1] The argument is part of his article *A nice derangement of epitaphs*, which continues the line of attack of his earlier article *Communication and convention.*[2] In the article, Davidson pursues two goals: he argues against conventionalists and he develops his alternative account. For the purposes of this chapter, I limit myself to a discussion of what I take to be Davidson's argument. Only in passing will I mention his positive account. Moreover, I won't discuss Davidson's argument from novel expressions (a hearer can understand novel expressions). But, there is an immediate problem: Davidson has never given a concise statement of what he takes to be his main argument.

Unsurprisingly, what has been concluded from Davidson's writing also varies. Some recent examples are Kemmerling (1993), Reimer (2004), Tiel-

[1] Davidson doesn't distinguish conventions from social norms as I do but his points apply equally to social norms. To streamline the discussion, I won't talk about social norms in this chapter. One should understand the points about conventions here as points about stable uses (*i.e.* conventions, social norms, and normative conventions) in general.

[2] See Davidson (2005) and Davidson (2001), respectively. Within this chapter, references by page alone refer to (Davidson 2005).

mann (2005), Horowitz (2007), and Lepore and Ludwig (2007). Of the six mentioned authors, all but Horowitz reject "the argument" but for different reasons. Kemmerling wants to defend a conventionalist account in which the notion of a public language is a necessary part. Tielmann defends an "idiolect first" position and thus might be expected to join Davidson's attack but nevertheless, he thinks that Davidson is wrong. Lepore and Ludwig provide a balanced discussion but reject some of the stronger claims of Davidson. Reimer disagrees with Davidson about what malapropisms mean. And finally, there is Horowitz who endorses Davidson's argument and also thinks that Davidson's positive view is correct. Since none of these authors give a concise statement of Davidson's argument, I will reconstruct the argument independently.[3]

Talking about "the argument" without giving some indication what it is, is not very satisfying. What is at stake? Suppose you and I are engaging in a conversation and at some point I utter the words:

(1) I take for granite that you will come to my housewarming party.

You don't protest and have understood my words as saying that:

(2) I take for granted that you will come to my housewarming party.

Without much ado, you corrected my linguistic mishap of uttering "granite" instead of "granted." Davidson's question is how we can explain the communicative success in cases of such linguistic mishaps, or "malapropisms"

[3] I recommend (Tielmann 2005:101–155) for a scholary assessment and (Lepore and Ludwig 2007:§17) for a discussion of important issues. Tielmann provides a reconstruction and a thorough and critical discussion of Davidson's article (without explicitly stating what I take Davidson's argument to be). It's by far the most complete discussion linking Davidson's article to other ideas he had. Lepore and Ludwig focus in chapter 17 on Davidson's claim about the reality of language and discuss not only the argument I discuss but also several others (but also without explicitly reconstructing Davidson's argument in normal form). I agree mostly with what they say, the difference being mostly in style. I was interested in a reconstruction of Davidson's argument. They are more interested in a systematic approach than in the details of Davidson's development of his argument. Kemmerling (1993) clarifies many underlying issues of Davidson's attack on conventionalists. He makes the communicative role of meanings explicit by discussing Davidson on the basis of a model of communication. He links this model to the so-called "standard account" of language and linguistic communication which is inherent in many conventionalist accounts, see (Kemmerling 1993:87–88). One of the merits this approach is that the model illuminates what the issues are (but Tielmann (2005:145–151) rightly disagrees with parts of the model).

as he calls them. According to Davidson, malapropisms are interesting since it turns out that (the notion of) a convention is neither necessary nor sufficient to explain successful communication in instances of linguistic communication with them. Hence, classical conventionalist accounts cannot explain communicative success in such cases. But thereby, conventionalists have to give up an important part of their project, namely that (the notion of) conventional meaning is part and parcel of an explanation of linguistic communication.

The reason why conventionalists can't explain communicative success in cases of malapropisms is this: A conventionalist about meaning claims that an expression's literal (or conventional) meaning is something learned in advance of actually using it. However, in cases of successful communication with a malapropism, it seems that what the speaker's utterance means is not the learned meaning of the words uttered. Nevertheless, hearers often understand such utterances in the speaker's intended meaning. Hence conventionalists cannot explain that hearers actually understand malapropisms.

Arguably, linguistic mishaps often happen to language users. Hence, malapropisms are theoretically important since one cannot just ignore these cases by claiming they are marginal. So, conventionalists either have to restrict the applicability of their accounts and provide an additional account to close the explanatory gap, or their accounts are wrong because they give the wrong results in important cases. I'll argue that the applicability of conventionalist accounts doesn't have to be restricted. Yet, an additional account is required to close the explanatory gap.

In short, we have a *prima facie* threat against conventionalists in general, not only against a particular account like Lewis'. But so far, it's not yet an argument. Hence, it will be of interest to see how Davidson proceeds to develop the threat into an argument.

3.1 Reconstruction

Let us begin with the end of Davidson's article and reconstruct how he ended up there. This is what I take to be the main conclusion of *A nice derangement of epitaphs*:

> We must give up the idea of a clearly defined shared structure which language-users acquire and then apply to cases. And we should try again to say how convention in any important sense is involved in

> language; or, as I think, we should give up the attempt to illuminate how we communicate by appeal to conventions.
>
> (Davidson 2005:107)

What Davidson means by a "clearly defined shared structure" is a system of rules which can be used to derive what linguistic expressions mean. His favorite way to express such a system is as a Tarskian truth theory. But, as far as I understand, nothing depends on the particular choice. Davidson's point is that language users do *not* apply such a clearly defined shared structure when they engage in linguistic communication. That is, linguistic communication is not like applying a code book where the speaker *encodes* what she wants to say as a message which the hearer then *decodes*. This picture seems to be inherent in many conventionalist accounts (at least, in Signaling Games and Actual Language Relation accounts).

The conclusion already indicates some ingredients of Davidson's argument: It's about instances of communication and about explaining the communicative success (or failure) in terms of linguistic competence. Apparently, conventionalists get it wrong when they make a claim about what the communicative role of conventions is.

To get some such argument for Davidson's denial off the ground, we need a claim what successful linguistic communication is, assumptions about the role of conventions in linguistic communication, and cases which convince us that conventions cannot have this role. Thus, we have a template of the structure of the argument. Davidson provides such an argument. Before I will state it in §3.2, I reconstruct its ingredients: A characterization of a notion of meaning tied to successful linguistic communication called "first meaning" (§3.1.1), its communicative role – also with respect to the conventionalist's claim (§3.1.2), and the troublesome phenomenon: malapropisms (§3.1.3).

3.1.1 First meaning

Davidson characterizes the role of linguistic conventions in communication by characterizing the role of conventional meaning in communication. To this end, he introduces two different notions of meaning which he calls "first meaning." One of them is used to capture what successful communication – *but not necessarily linguistic communication* – consists in. The other notion applies only to linguistic communication, successful or not. It is supposed to be an explicit characterization of so-called “literal meaning” or, equivalently in this context, “conventional meaning.” I will refer to the former notion as “generic first meaning” and to the latter as “specific first meaning.”

Generic first meaning Unofficially, “generic first meaning” is so called because it’s the first meaning a hearer arrives at in the interpretation of an utterance. The official characterization is as follows:

> The concept applies to words and sentences as uttered by a particular speaker on a particular occasion. [...]. Roughly speaking, first meaning comes first in the order of interpretation. [...] But “the order of interpretation” is not at all clear. [...] And of course it often happens that we can descry the literal meaning of a word or phrase by first appreciating what the speaker was getting at. A better way to distinguish first meaning is through the intentions of the speaker. The intentions with which an act is performed are usually unambiguously ordered by the relation of means to ends. [...] The order established here by “by” can be reversed by using the phrase “in order to.” In the “in order to” sequence, first meaning is the first meaning referred to. (“With the intention of” with “ing” added to the verb does as well.) (Davidson 2005:91–92)

What Davidson considers here are different ways of expressing a means-to-end relation, *i.e.* in the following two ways: (i) “agent S did γ_n by doing γ_{n-1}. ... S did γ_2 by γ_1” and (ii) “agent S did γ_1 with the intention of γ_2-ing. ... S did γ_{n-1} with the intention of γ_n-ing.”

When the means-to-end relation is about communication, then S is a speaker and her action γ_1 is the uttering u of some expression. Davidson’s claim is that the intentions of a speaker are usually ordered unambiguously. Below, I will use the “with the intention of”-formulation because it mentions *intentions* explicitly.

On the basis of this means-to-end relation, Davidson illustrates generic first meaning as follows:

> Suppose Diogenes utters the words “I would have you stand from between me and the sun” (or their Greek equivalent) with the intention of uttering words that will be interpreted by Alexander as true if and only if Diogenes would have him stand from between Diogenes and the sun, and this with the intention of getting Alexander to move from between him and the sun, and this with the intention of leaving a good anecdote to posterity. Of course these are not the only intentions

> involved; there will also be the Gricean intentions to achieve certain ends through Alexander's recognition of some of the intentions involved. Diogenes' intention to be interpreted in a certain way requires such a self-referring intention, as does his intention to ask Alexander to move. In general, the first intention in the sequence to require this feature specifies the first meaning. (Davidson 2005:92)

So, when Diogenes utters these words with these complex intentions ordered by a means-to-end relation, the first intention in this order determines the generic first meaning of the words uttered on that particular occasion of use.

With this in mind, we can characterize *generic first meaning* as follows: It is a concept that applies to expressions on a particular occasion of use. The first meaning of an expression on a particular occasion of use is what the speaker intends the expression's words to mean for a certain audience on a particular occasion of use.[4] More precisely, it's the speaker's *first* communicative intention in the means-to-end relation which determines what the speaker's words mean. According to Davidson (p. 92), it should be the first intention in this relation since a speaker might also have further Gricean intentions to achieve some ends by the hearer's recognizing some of the speaker's intentions. But these further intentions should not determine the expression's generic first meaning. (I'll return to this point later.)

As Davidson observes, the application of generic first meaning is not restricted to words (see p. 93). It can also be applied to cases of non-linguistic communication. Hence it wouldn't be different from speaker-meaning in this respect.[5]

But Davidson wanted to uphold the distinction between speaker-meaning and literal meaning (see p. 91). Literal meaning should apply to linguistic expressions and only to them. Hence, Davidson needs another notion of meaning to have something akin to so-called "literal meaning."

[4] On p. 92, Davidson offers another characterization of first meaning in terms of the hearer's knowledge or the abilities she must have if she is to interpret a speaker. I'll focus on the first characterization.

[5] More precisely, I agree with the judgment of Kemmerling (1993:99 fn. 19): "It is much more like a hybrid between what Grice calls utterer's occasion meaning and what he calls applied timeless sentence meaning" in the speaker's idiolect.

Specific first meaning As a candidate for a notion of literal meaning, Davidson introduces the notion of *specific first meaning.*[6] According to Davidson, the notion is supposed to capture the conventionalists' commitments about literal (or conventional) meaning (p. 91). An expression's specific first meaning is its generic first meaning which satisfies the following three conditions (p. 93):[7]

FM1. *First meaning is systematic.* A competent speaker or interpreter is able to interpret utterances, his own or those of others on the basis of the semantic properties of the parts, or words, in the utterance, and the structure of the utterance. For this to be possible, there must be systematic relations between the meanings of utterances.

FM2. *First meanings are shared.* For speaker and interpreter to communicate successfully and regularly, they must share a method of interpretation of the sort described in FM1.

FM3. *First meanings are governed by learned conventions or regularities.* The systematic knowledge or competence of the speaker or interpreter is learned in advance of occasions of interpretation and is conventional in character.

Davidson was aware that "there are difficulties with these conditions" (p. 93), but let us focus on the important points. First, the specific notion is one which applies to linguistic communication involving words, phrases, and sentences – and only to them. Non-linguistic communication, like Emmanuel Rahm's mailing a rotting fish to someone hated, is thereby ruled out.[8] This is the effect of FM1. Second, a necessary condition for successful linguistic communication is that generic first meaning is shared between the speaker and the hearer. Third, an expression's specific first meaning is determined before it is actually interpreted on a particular occasion of its use (FM3). Generic first meaning also determines a meaning on a particular occasion of use. As I will argue, this leads to problems (§3.1.4). Fourth,

[6] In later writing, Davidson confesses that he sometimes used the expression "first meaning" when he really meant "literal meaning": "I confess that having explained what I meant, I have sometimes allowed myself to substitute the *phrase* 'literal meaning' for 'first meaning.'" (Davidson 1993:118) Thus, my terminological separation of generic first meaning and specific first meaning (aka "literal meaning") has the merit of avoiding possible misunderstandings.

[7] Except for the labels FM1–3, the conditions are literal quotations from Davidson's article.

[8] See (Kintisch 30.11.1999). Rahm has been appointed by Obama in 2008 as his new White House Chief of Staff.

specific first meaning is conventional (FM3).

We can now state a first ingredient of Davidson's argument. According to conventionalist accounts, the literal meaning of an expression on a particular occasion of use is its *conventional meaning* (when relativized to a particular context, as Reimer (2004:320) points out). We can express this claim as follows:

I1. According to conventionalist accounts, the literal meaning of an expression on a particular occasion of use is its *specific first meaning.*

Ingredient I1 is an identity claim about literal meaning *being specific first meaning* which he puts in the mouth of conventionalists. The claim turns out to be false and requires a slight revision (*cf.* §3.1.4).

3.1.2 The role of first meaning in communication

Davidson's characterization of specific first meaning entails a claim about its communicative role in the context of successful linguistic communication. The claim is crucial for his argument. I take Davidson's condition FM2 of specific first meaning to express such a claim about the communicative role of first meaning. But more clear is perhaps the next quotation:

> Because a speaker necessarily intends first meaning to be grasped by his audience, and it is grasped if communication succeeds, we lose nothing in the investigation of first meaning if we concentrate on the knowledge or ability a hearer must have if he is to interpret a speaker. (Davidson 2005:92)

This quotation occurs in the context of Davidson's move from the speaker's linguistic knowledge to the hearer's. I want to focus on that part of the quotation which says, roughly, that *if communication succeeds, then first meaning is grasped.* Since this quotation occurs in the text before Davidson introduces what I call "specific first meaning," "first meaning" refers to *generic first meaning.* So, paraphrasing the condition in my terminology, we can express the second ingredient of Davidson's argument as follows:

I2. A necessary condition for successful linguistic communication is that the hearer understands the expression the speaker uttered in its generic first meaning.

An issue with the quotation above is that what I've paraphrased is embedded in the clause beginning with "Because a speaker necessarily intends first meaning to be grasped by his audience."[9] But other quotations in Davidson's article support my interpretation.[10]

3.1.3 The cases: Malapropisms

The phenomenon of malapropisms consists in a certain kind of linguistic mishap, an "incorrect" and possibly novel use of an expression, which is such that linguistic communication nevertheless often succeeds, that is, hearers often successfully understand the speaker's words. *Linguistic mishaps* should be contrasted with *factual errors* (*e.g.* when someone is misinformed and, because of that, says something which is not true). But Davidson does not give an explicit definition of the phenomenon.[11] Thus, the best we can do is to consider some examples and point out some of their important properties.[12]

The first example is from Sheridan's play *The rivals* in which Mrs. Malaprop says that something is "a nice derangement of epitaphs" but she intended her words to mean *that it is a nice arrangement of epithets* (Sheridan 1775). Another example is from Marga Reimer who reports:

> One of my students insists that "for all intensive purposes" we are at war with al Qaeda. (Reimer 2004:317)

Surely, the student wanted to say that *for all intents and purposes* we are at war with al Qaeda. Plausibly, this was a *slip of the tongue.* But

[9] There are other issues with this quotation which are discussed by Tielmann (2005:115 ff.).

[10] On p. 93, Davidson writes: "[I]f the speaker is understood he has been interpreted as he intended to be interpreted." We can restate the antecedent of the quotation "If the speaker is understood" as "If linguistic communication is successful." For the quotation occurs in the context of thinking about linguistic communication and "to be understood" just seems to mean that linguistic communication is successful. The consequent of the quotation plausibly expresses that the meaning the hearer grasped when she understood the speaker's words is the words' generic first meaning. If so, also this second quotation supports my interpretation. A third quotation supporting my interpretation occurs in the context of Davidson's positive proposal in terms of prior and passing theories on p. 102. There, Davidson proposes a success condition for linguistic communication that is stronger than I2. For the argument, I2 is sufficient.

[11] On pp. 89–91 and pp. 94 ff. he provides a lengthy characterization.

[12] But it's unclear whether what has been called "malapropisms" is a homogeneous phenomenon; *cf.* also (Schulte 1993). George (1990:278 ff.) seems to be more optimistic and distinguishes between different kinds of malapropism.

malapropisms needn't be unwittingly produced. The examples Davidson cites from Mark Singer about Goodman Ace illustrate this point nicely. Ace was a radio sitcom writer who often talked the way Singer wrote:

> Rather than take for granite that Ace talks straight, a listener must be on guard for an occasional entre nous and me . . . or a long face no see. In a roustabout way, he will maneuver until he selects the ideal phrase for the situation, hitting the nail right on the thumb.
>
> (Davidson 2005:89)

Davidson's point is that hearers can – and actually do – successfully understand malapropisms. This observation is the argument's third ingredient:

I3. There are cases (occasions) of successful linguistic communication with malapropisms.

Davidson's twist to his observation is that he claims that these cases are cases of successful *linguistic* communication where what the hearer understands is the words' *literal meaning* (p. 102). This claim is best understood as a theoretical stipulation about "literal meaning". If I understand Davidson's enterprise correctly, then the stipulation is justified by the theoretical role literal meanings are supposed to play in a theory of linguistic communication: it's a "notion of what words, when spoken in context, mean" (p. 91). Moreover, it's that which a hearer grasps in case of successful communication (I2).

According to Davidson's characterization of the conventionalist, the words' literal meaning is their *specific first meaning.* This is the fourth ingredient:

I4. According to conventionalist accounts, in case of successful linguistic communication with a malapropism, a hearer understands the malapropism the speaker uttered in its specific first meaning.

3.1.4 Towards the argument

The problem for the conventionalist is now this. Consider the utterance of (3):

(3) "Something is a nice derangement of epitaphs."

a. Generic first meaning: *that something is* <u>*a nice arrangement of epithets*</u>

b. contextually-relativized conventional meaning: *that something is a nice derangement of epitaphs*

Let us assume that on this occasion, linguistic communication is successful with the malapropism (I3). According to the success condition (I2), linguistic communication is only successful if the hearer understands the malapropism in its generic first meaning.

For the sake of the argument, let us agree that the generic first meaning on that particular occasion of use is (3-a).[13] The malapropism's conventional meaning is learned in advance of its use. It amounts roughly to what is said by it (in the everyday sense of "what is said" and not in one of the theoretically loaded senses). Let us agree that the malapropism's contextually-relativized conventional meaning is (3-b). This brings us to a general observation:

I5. In cases of successful communication with malapropisms, the malapropism's generic first meaning is not identical to its contextually-relativized conventional meaning.

According to Davidson, successful communication in such cases cannot be explained, as the conventionalist claims (I4), on the basis of the malapropism in its contextually-relativized conventional meaning. The hearer has to understand the malapropism in its generic first meaning for communication to be successful. Since Davidson tacitly makes the following additional assumption,[14] the conventionalist has a problem:

I6. In cases of successful communication with malapropisms, if a hearer understands an expression a speaker uttered in its generic first meaning, then she doesn't understand it in its (contextually-relativized) conventional meaning (and *vice versa*).

This seems to be a fair restatement of what Davidson declares to be the problem:[15]

[13] This is by no means obvious, as Kemmerling (1993:101) points out since what the speaker-meaning of an utterance is depends on many highly contextual factors of the case and on background assumptions about force and content.

[14] Davidson makes this assumption by defining generic first meaning in terms of the *first* communicative intention (*cf.* §3.1.1). Thereby, understanding an expression in one meaning by means of understanding it in another meaning is ruled out.

[15] A hearer's *passing theory* is the semantic theory she actually uses to interpret the speaker's utterance (*cf.* p. 101). The interpretation the theory entails for the uttered expression

> Stated more broadly now, the problem is this: what interpreter and speaker share, to the extent that communication succeeds, is not learned and so is not a language governed by rules or conventions known to speaker and interpreter in advance; but what the speaker and interpreter know in advance is not (necessarily) shared, and so is not a language governed by shared rules or conventions. What is shared is, as before, the passing theory; what is given in advance is the prior theory, or anything on which it may in turn be based.
> (Davidson 2005:105–106)

However, we cannot yet state Davidson's argument. For I presented the problem in terms of *conventional meaning* and not, as Davidson does, in terms of *specific first meaning.* Doing so uncovers a problem in Davidson's rendering of the conventionalist's position. For in cases of successful communication with malapropisms, there is no generic first meaning that satisfies conditions FM1–3 required for specific first meaning.

The reason is this. Specific first meaning is defined as generic first meaning satisfying conditions FM1–3. I've already pointed out that generic first meaning is determined by the first communicative intention in the chain of intentions the speaker has on particular occasions. However, it seems that conditions FM1–3 determine *on their own* something akin to *conventional meaning*, independently of generic first meaning. In particular, there is condition FM3: "First meanings are governed by learned conventions or regularities." Thus, a plausible reading of FM3 is that specific first meanings of expressions are *determined* by learned conventions or regularities.

This leads to an inconsistency if we unpack the definition of specific first meaning. The reason is that it has two meaning-determining components which might not coincide. Consider again the example of Mrs Malaprop. The example illustrates that the malapropisms' contextually-relativized conventional meaning (3-b) is not identical to its generic first meaning (3-a). By definition, specific first meaning is generic first meaning *and* conventional meaning. Hence, specific first meaning is not defined because there is no way to satisfy its definition as being generic first meaning satisfying conditions FM1–3. Either conditions FM1–3 are satisfied or the generic-first-meaning

needn't coincide with the interpretation the hearer has learned before having started to interpret the utterance. A passing theory is a theory "geared to the occasion" (*ibid.*). For example, the learned interpretation for "take for granite" might be *take for granite* – which amounts to nonsense – but the typical hearer will likely interpret an utterance of it as *take for granted.* This latter interpretation is then based on the passing theory.

part is satisfied. This is so whenever the two meaning components do not coincide, as in cases of malapropisms.

This outcome suggests strongly to me that Davidson's statement of the conventionalist position is not faithful to the conventionalist. For conventionalists do not claim that malapropisms are meaningless. They just claim that the meaning is a different one than Davidson proposes. But for the problem Davidson observed, it is not crucial that we state it in terms of *specific first meaning*. Hence, I propose to reconstruct Davidson's argument in terms of *conventional meaning* (thereby we give up I1 and I4).

3.2 The argument

The ingredients I1–4 have served us well so far. But it will be helpful to formulate the premises differently to state Davidson's argument:

P1. According to conventionalist accounts, a hearer understands the malapropism the speaker uttered in its (contextually-relativized) conventional meaning.

P2. A necessary condition for successful linguistic communication is that the hearer understands the expression the speaker uttered in its generic first meaning.

P3. There are cases (occasions) of successful linguistic communication with malapropisms.

C1. In such a case, the hearer understands the malapropism the speaker uttered in its generic first meaning.

P4. In such a case, (contextually-relativized) conventional meaning is not identical to generic first meaning.

P5. In such a case, if a hearer understands an expression a speaker uttered in its generic first meaning, then she doesn't understand it in its (contextually-relativized) conventional meaning (and *vice versa*).

C2. Conventionalist accounts cannot be correct.

P1 characterizes the relevant conventionalist claim about meaning. It has the status of an assumption to derive a contradiction. P2 and P3 are I2 and I3 from above. C1 is a consequence about what is the case when a speaker successfully (linguistically) communicates with a malapropism. C1 follows from P2 and P3. Finally, P4 and P5 are I5 and I6 from above. This entails a contradiction. For by C1 the hearer understands the malapropism's in its generic first meaning. Hence, by C1 and P5, she understands the malapropism not in its conventional meaning. But according to the conventionalist, she does (P1). Consequently, we reject the assumption P1 by *reductio ad absurdum*. Since P1 is part and parcel of conventionalist accounts, they cannot be correct.

So, it is clear how Davidson arrived at his conclusion that conventionalists either have to come up with a better proposal or, as he favors, to give up conventionalist accounts altogether. For if Davidson's argument is sound, then it shows that conventional meanings of expressions cannot be identical to *what the speaker intends the hearer to understand*. But this is, according to Davidson, constitutive for literal meaning and required for explaining communicative success. Hence, conventional meaning can neither be (a kind of) literal meaning nor explain communicative success.

3.3 Evaluation of the argument

Davidson's argument is valid and its conclusion removes the possibility of an adequate conventionalist account. I'll argue that it is not sound. First, let me observe that independently of Davidson's argument, conventions are not required for successful communication. A speaker can get across what she wants whether or not conventions exist. This has, however, not so much to do with *linguistic* communication but with communication in general: To get across what one means, it can be sufficient that that hearers recognize the speakers' intention. This much is uncontested.

It seems to me equally uncontested that the intention-recognition alone isn't sufficient to explain linguistic communication. For how could one explain that language users can reliably and easily communicate complex and unexpected contents by uttering expressions? To do so, another explanation is required. For the ordinary cases of "literal communication," conventionalists offer a convincing explanation. Without conventions it seems very difficult to offer a plausible explanation of this feat. So, one shouldn't deny that conventions play a role in ordinary cases.

The critical cases are malapropisms. Successful communication in these cases is not explained by the hearer's understanding the malapropism in its conventional meaning. Hence, a more modest conclusion from Davidson's argument would be that conventionalist accounts cannot explain successful communication in these cases.

But even this conclusion is too strong. What goes wrong is, I think, the assumption of P5. According to it, if a hearer understands an expression in one meaning, then she doesn't understand it in another meaning. But a malapropism can also be understood in its generic first meaning if the hearer bases her understanding on the malapropism's conventional meaning.[16] This amounts to something like an extended Gricean strategy to explain non-standard meanings: the hearer notices that the conventional meaning makes no sense and, possibly guided by contextual clues and an inference to the best explanation, she concludes that she should understand the speaker's words in their generic first meaning.

This explanation seems plausible. First, it's coherent with usual Gricean explanations for implicatures and the like. Second, successful communication with malapropisms seems to depend on background information about phonetic, syntactic, semantic, and pragmatic similarities of utterances of expressions. For example in the "take for granite" case, there is a phonetic similarity between "granite" and "granted." There is also syntactic similarity between "take for granite" and "take for granted." A conventionalist can explain why these similarities exist and why they are – in some sense – *known* to the dialog partners: They supervene globally on the conventional use of the expressions in the language. Without assuming conventions, such an explanation is not readily available.

For these reasons P5 should be rejected. But then, the problematic conclusion does not follow.

3.4 Summary

Contra Davidson, there are good reasons not to accept the conclusion that is problematic for the conventionalist project. Nevertheless, Davidson has a point when he draws attention to malapropisms: To explain successful communication with them, a conventionalist would need an additional explanation.

But if my characterization of the conventionalist project is correct, then

[16] To make this work, we have to define "generic first meaning" in another way. For it's then not necessarily the meaning fixed by the *first* communicative intention but maybe by another intention the speaker had.

conventionalist accounts should not be understood as theories of linguistic communication. A conventionalist account primarily makes claims about the determination of expressions' meanings. In doing so, some commitments about language use and linguistic communication are incurred. But they are relatively unspecific. In particular, it's compatible with a conventionalist account that novel expressions are used in whichever way. A conventionalist has nothing to say about such uses. A conventionalist's claim is conditional: If an expression has a conventional use, then it has a certain meaning. Problematic are thus cases where expressions have established uses but are nevertheless used non-conformingly on particular occasions. Here, it's a matter of proportion: How often do such cases occur? Are they recognized as deviations? Only if such cases occur quite often and are not recognized as deviations, the conventionalist has a problem. For then it's unclear whether she is able to assign meanings to these expressions.

Hence, we shouldn't dismiss Davidson's insights too easily. As we will see in chapter 6, in the first version of his account, Lewis ignored *Davidsonian uses* (*i.e.* mistakes, non-literal, and novel uses of expressions).[17] The initial omission was a disfavor to the conventionalist project.

Moreover, even if Davidon's argument fails, the second horn of its conclusion – his advise to "try again to say how convention in any important sense is involved in language" is very much to the point: Conventionalists need to commit themselves to a pragmatic theory (i) to define what they mean be "literal meaning" (see §1.1.1) and (ii) to indicate how Davidsonian uses are explained.

Another lesson we should learn from Davidson (*cf.* also (Savigny 1985) and (Sperber and Wilson 1995)) is that the linguistic behavior is flexible and allows for variation. The challenge for a conventionalist (as I understand it) is then to make elbow room for variations in the linguistic behaviors of communicating language users.

[17] In Lewis' *Convention* (Lewis 2002) words like "(linguistic) mistake," "non-literal" and its cognates do not even occur. I used Amazon OnlineReader to establish this claim. In his later article *Languages and Language* (Lewis 1975:395 ff.), Lewis' account is extended to Davidsonian uses. I will return to this topic in chapter 6.

Chapter 4

Lewis conventions

> Conventions are regularities in action, or in action and belief, which are arbitrary but perpetuate themselves because they serve some sort of common interest.
>
> *Languages and language*
> David Lewis

Reassured by the defense of conventionalism in the last chapter, it is time to discuss what has become the standard analysis of conventions, namely the one Lewis started to develop in the 1960s, succinctly summarized above. After forty more years of analytic philosophy, Lewis' analysis is still one of the landmark pieces of analytic philosophy. Both exposition as well as analytic insight make it a marvel. Geniality notwithstanding, the analysis has been found by many to be utterly implausible. But more often than not, the harsh judgment is based on misunderstandings. In this chapter, my concern is thus to present and motivate Lewis' analysis and to discuss some objections.

The analysis is put to use subsequently in my discussion of conventionalist theories of two research paradigms, namely Signaling Games and Actual Language Relations.

In §4.1 I present Lewis' account. In §4.2 I provide its game theoretical foundations and a simpler analysis considered by Lewis. The account is evaluated in §4.3. The chapter ends with a summary in §4.4.

4.1 Lewis' account

Lewis set out to provide a definite analysis of the notion of convention in a series of writings to which notably *Convention* (Lewis 2002) belongs. To fix the notion, Lewis presents a list of central cases of social situations in which people coordinate. Lewis invites us to call these central cases "conventions." Among the central cases are the ones I used to introduce the pre-theoretic notion of a convention (§1.2). Thereby, Lewis' notion (unsurprisingly) satisfies my pre-theoretic characterization. To remind you of two of Lewis' examples:

(1) **Rowing a boat**: You and I sit in a boat and want it to glide in one direction while speed is not important to us. As long as we both row with the appropriate frequency and strength, we will satisfy our interest. There are many combinations of individual frequencies and strengths which are in this sense equally good. We both row in a way satisfying the shared goal.

(2) **Driving right**: Car drivers share an interest in efficient and safe conduct on the streets they drive. Among the many ways they could behave when they drive, two of them are particularly simple: Crossing drivers drive on their respective right (or left – the second option). By doing so, they further their common interest. They drive on their respective right.

Cases like these exhibit several interesting properties. First, there are *at least two agents* involved who might be drawn from a population as (2) illustrates. Second, every agent has *at least two available ways of acting* in the situation. In (1) each individual has a range of actions at her disposal which are different ways of rowing (individuated by strength, frequency, direction, ...). Third, in a particular situation of the relevant type, the outcome of the action of each agent *depends on the actions of the other involved agents.* In (1) one's rowing either results in gliding straight or in circles. Fourth, all involved agents *have a common interest in attaining a certain goal*, *e.g.* gliding straight. Fifth, there are *at least two ways to attain the goal.* Both are in a sense equally desirable: when attaining the goal in one way, each prefers doing one's part of it to not doing it; but if the others do their part of another way, then one prefers to do one's part of that way.

If a situation satisfies these five conditions, then each involved party faces a problem: There is more than one way to attain the goal. Each requires one to act differently. But only if one coordinates with the others, is the goal attained. Since this is the case for every agent, it's a difficult problem to solve.

Let us say that such a situation has the "structure of a coordination problem" or simply, that it *is* one. Typically, such situations are not one-off problems. They reoccur. Lewis' proposal is to conceive of conventions as regularities bringing about coordination in such situations.

Lewis' two definitions of a convention There are two definitions of a convention which Lewis has endorsed at some point. The first definition was published in *Convention* (Lewis 2002), the second in *Languages and language* (Lewis 1975). Both are of interest. The former helps to explain certain features of Lewis' Signaling Games theory and why the latter definition is as it is. The first definition is:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P,
>
> 1. almost everyone conforms to R;
> 2. almost everyone expects everyone else to conform to R;
> 3. almost everyone has approximately the same preferences regarding all possible combinations of actions;
> 4. almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R;
> 5. almost everyone would prefer that any one more conform to R', on condition that almost everyone conform to R', where R' is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R' and R. (Lewis 2002:78)

Example: Consider the population of Dutch car drivers (P). Among them is a regularity to drive right on public roads (R). When they drive on public roads (S), then they drive right (clause 1). They expect the others to drive right (clause 2). They have approximately the same preferences regarding *all* combinations of actions in the coordination problem (*e.g.* both driving left; both driving right; one driving left, the other right; one driving right, the other left; and more complex combinations). With respect to driving, they prefer to drive right, on condition that others do, but dislike to drive right, on condition that the others drive left. They also prefer to drive left, on condition that the others do, but they dislike to drive left, on condition that the others drive right (clause 3).[1] They prefer that the others also drive right, on condition that they drive right (clause 4). They would prefer to drive left, on condition that the others did so as well (clause 5). Finally, whenever they drive on public roads, they know that the clauses 1 to 5 are satisfied, they know that the others know, they know that the others know that they know, and so on (the common-knowledge assumption).[2]

That does not sound too implausible. But the definition has problems which led Lewis to propose a second. Let me mention *some* of the issues:

First, if we want to count regularities of linguistic communication as conventions, then the notion of a regularity in *action* is too narrow. As Bennett (1973:150) pointed out, plausibly such regularities consist in (something like) a speaker's meaning something and a hearer's understanding it. But neither the speaker's meaning something nor the hearer's understanding are *actions*. A wider notion of regularity is needed. Bennett proposed *regularities in action and belief*; Lewis (1975:11 ff.) accepted. But there is no good reason to limit the regularities to action and belief – every mental state for which one can have reasons will do; so I suggest to generalize the proposal to regularities in *action and attitude*.[3]

Second, the talk about "expectations" has not really been made precise and invites misunderstandings of it as a normative concept (as in: "I expect (=demand) you to conform").[4]

Third, according to the first definition, almost all members of a possibly large population must have "approximately" identical preferences. This is much too demanding and luckily not required. It suffices that there is enough common interest to realize a certain goal.

Lewis improved on the earlier definition in *Languages and language*, yielding what I call his "official" definition:

[1] Conditional preferences are neither *conditionals about preferences* ("If p, then I prefer a to b") nor *preferences among conditionals* ("I prefer $a \to b$ to $c \to d$"). Rather, to prefer a to b, conditionally on c is to prefer the combination of c and a to the combination of c and b; to prefer a on condition that p is to prefer a to non-a, conditionally on p. *Cf.* (Lewis 1976:117 ff.) on conditional preferences, arguing against (Jamieson 1975:75 ff.).

[2] Strictly speaking, additional assumptions are required, see §4.3.1.

[3] Thereby the conception of game theory is not (strictly) behavioristic anymore since mental activities can be part of the regularities. However, the relevant mental activities are expressed in systematic ways; hence, the regularities still have an important behavioral part.

[4] *E.g.* Itkonen (2008) is guilty of such a misunderstanding (§2.2.2). Lewis (2002:97) is very clear to point out that his definition is not defined in normative terms.

> [A] regularity R, in action or in action and belief, is a *convention* in a population P if and only if, within P, the following six conditions hold. (Or at least they almost hold. A few exceptions to the "everyone"s can be tolerated.)
>
> 1. Everyone conforms to R.
> 2. Everyone believes that the others conform to R.
> 3. This belief that the others conform to R gives everyone a good and decisive reason to conform to R himself. [...]
> 4. There is a general preference for general conformity to R rather than slightly-less-than-general-conformity – in particular, rather than the conformity by all but any one. [...]
> 5. R is not the only possible regularity meeting the last two conditions. There is at least one alternative R' such that the belief that the others conformed to R' would give everyone a good and decisive practical or epistemic reason to conform to R' likewise; such that there is a general preference for general conformity to R' rather than slightly-less-than-general conformity to R'; and such that there is normally no way of conforming to R and R' both. [...]
> 6. Finally, the various facts listed in the conditions [1.] to [5.] are matters of *common* (or *mutual*) knowledge: they are known to everyone, it is known to everyone that they are known to everyone, and so on. The knowledge mentioned here may be merely potential: knowledge that would be available if one bothered to think hard enough. [...] (Lewis 1975:5–6)

The two most important changes are, arguably, the relaxation with regard to the regularities and the preference structures. The regularities can now also in action and belief (or as I propose: in action attitudes) and not only be in action. The agents' preferences needn't coincide anymore; it suffices that there are two combinations of actions (and attitudes) that are generally preferred (clauses 4 and 5 of the second definition).

By these relaxations, the second definition is more inclusive than the first.[5] Now there can be a convention in a population whose members have mostly opposing preferences but which coincide with respect to two

[5] There is one proviso: The second definition might also be more restrictive as to what counts as a convention. In the first definition, what is called "common knowledge" amounts to *common reason to believe*. In contrast, common knowledge is according to the second definition a kind of *knowledge*; see my discussion of common knowledge below in §4.3.1.

combinations of actions.

Not having introduced the concepts the two definitions use – *regularities*, *preferences*, *alternatives*, it remains unclear what the conditions amount to. In particular, what is a "general preference for general conformity" (condition 4)? What is a "good and decisive reason"? And what is it to give someone such a reason? I turn to these questions in the next two sections.

4.2 Theoretical foundations: Game theory

Lewis' account has its roots in (classical) game theory, wherein the notions of preference, alternative, and common knowledge are explicated.[6]

Moreover, using game theory, Lewis suggested and rejected a simpler definition of a convention than the two we've considered above. I think it can be rehabilitated. Also Signaling Games make use of games. For these reasons, I introduce the basics of game theory with a focus on coordination problems and reconstruct the simpler definition.

A central notion in game theory is the notion of a strategic game. By means of a special class of strategic games, so-called "coordination games," the notion of a coordination problem can be formalized. A strategic game is defined in terms of *agents*, *strategies*, and *preferences* over outcomes of possible strategy combinations. An example of a strategic game is the following version of the so-called "Battle of Sexes" ("BoS"):

(3) Judith and Marc individually deliberate about what to do. They want to have lunch together and join their respective friends who eat at different places. They have to choose where to go. Judith prefers joining her friends with him to joining his friends with him. Marc has it the other way round. They don't care about whom they join, if they don't have lunch together.

What should you do if you were in exactly the same situation as one of them? If they are rational agents, which course of action is expected? What is it

[6] Lewis' account is inspired by the work of Schelling (2005). By "(classical) game theory" I mean a version of game theory using a Savage-style decision theory with von Neumann/Morgenstern utility functions and an epistemic equilibrium interpretation, see (Neumann and Morgenstern 1953; Nash 1952; Savage 1972). Since Lewis' account uses game theory, it also faces the foundational issues of game theory; I won't go into this debate. But see (Resnik 2002) for a criticism of decision theory and (Alexander 2009:§3) for a criticism of classical game theory.

to act rationally anyway? Game theory answers questions of these sorts. The first question relates to a *normative interpretation* of game theory. So understood, the task of game theory is to establish standards of rational agency. Having a standard allows us to evaluate cases of agency with respect to their rationality – their "rational goodness" or "recommendability." The second question relates to an *explanatory interpretation* of game theory. If we expect that the agents in question are rational agents, then we can explain their behavior and make predictions about their future behavior. The third question relates to a *conceptual interpretation.* According to it, the role of (decision and) game theory is to explicate notions like desirabilities and subjective probabilities.

I will assume this last interpretation. Lewis endorsed it for decision theory (Schwarz 2009:16 ff.) and plausibly also for game theory. Moreover, I only consider two-player games without cooperation (*i.e.* the agents cannot make binding agreements) since this is all we need to study common cases of linguistic communication.[7]

Let us now develop the required theory by answering the following questions: (i) What is it to answer one's desires? (ii) What is a game and what is it to solve one? (iii) What are coordination games? (iv) What is common knowledge?

4.2.1 Decision theory

Decision theory studies decision problems. Decision problems are abstract representations of scenarios in which agents can choose what to do – or "can answer their desires." Agents are assumed to (i) be able to act, (ii) to have desires, and (iii) to deem events more or less likely. Their behavioral repertoire includes different kinds of behaviors which I will call, suiting the context, "action," "strategy," "contingency plan," or simply "behavior." Actions have consequences, or as one says in game theory, they lead to *outcomes.* The outcome an action leads to depends on the state the world

[7] Arguably, cooperation in the sense of game theory requires a meaningful language to reach agreements (and something like coinciding interests or vast sanctioning power). Hence cooperation should not be assumed. Otherwise, an explanatory circle would result if meaning in the agents' language is explained by means of another meaningful language (§5.3.3). If one wanted to use cooperative game theory, then one should disallow communication and only allow agreements reached by silently reflecting upon the strategic situation.

is in. Agents deem such states more or less likely. An outcome can be more or less desired or, synonymously, *preferred* by an agent in comparison to other outcomes. An agent's preferences can vary with place and time. For simplicity, I will treat them as constant. In general, we assume that an agent either *prefers* an outcome to another or is *indifferent* between them. An agent is indifferent between two outcomes iff she equally wants or would want that the respective other outcome obtains. In the BoS-example (3), if Judith has lunch alone with her friends, then she would equally want to have lunch alone with Marc's friends, and *vice versa*.

In a decision problem, some actions can be better than others; some are in a certain sense even optimal. Decision theory explains an action's optimality in terms of its expected utilities. They depend on the desirability and the subjective probability[8] of action's outcomes. To this end, a series of definitions is provided for (i) preferences over outcomes, (ii) expected utilities of actions, and (iii) a principle stating which action is optimal ("maximize expected utility!").

From preferences to expected utilities States of preferring outcomes to outcomes or being indifferent between them can be represented by a "preference relation". A *preference relation* is a (non-strict) weak order over the set of outcomes O (*i.e.* a binary relation on outcomes that is transitive and complete and thus reflexive).[9] Ordinal utilities u are numerical representations of preference relations. An *ordinal utility* u_o is a function from outcomes O into reals $\mathbb{R}$ such that for every two outcomes $o_1, o_2 \in O$: (i) $u_o(o_1) = u_o(o_2)$ iff the agent is indifferent between o_1 and o_2. (ii) $u_o(o_1) > u_o(o_2)$ iff the agent prefers o_1 to o_2. Ordinal utilities u and u' are equivalent iff for all outcomes $o_1, o_2 \in O$: $u_o(o_1) = u_o(o_2)$ iff $u'_o(o_1) = u'_o(o_2)$ and $u_o(o_1) > u_o(o_2)$ iff $u'_o(o_1) > u'_o(o_2)$. Ordinal utilities are unique up to order-preserving transformations. Hence, only the order but not the differences between utility values are meaningful.

Comparative strength of preferences can be represented by cardinal utilities.[10] A *cardinal utility* u is a function from outcomes O into reals $\mathbb{R}$

[8] A (subjective or objective) probability p is a function from the set of states T into $\mathbb{R}$ such that (i) $\forall t \in T : 0 \leq p(t) \leq 1$ and (ii) $\sum_{t \in T} p(t) = 1$.

[9] A binary relation $R \subseteq O \times O$ is called *complete* iff for all $a, b \in O$: aRb or bRa. This means that all elements in O are comparable. Since preference relations are non-strict, it's not implied that they are linear (*i.e.* that no two elements in O can be equally good).

[10] A discussion of the construction can be found in (Jeffrey 1990:41–58) where an alternative

satisfying conditions (i) and (ii) above and (iii) for outcomes $o_1, o_2, o_3 \in O$: if $u(o_1) > u(o_2) > u(o_3)$ with the differences $d_{12} = u(o_1) - u(o_2)$ and $d_{13} = u(o_1) - u(o_3)$, then $d_{13} = d \times d_{12}$ iff the agent's preference of o_1 to o_3 is to the degree d stronger than her preference of o_1 to o_2. Cardinal utilities u and u' are equivalent iff there is a positive m and b such that for all outcomes $o \in O$: $u(o) = m \times u'(o) + b$. Cardinal utilities are unique up to positive linear transformations. Hence, also the comparative differences between utility values are meaningful.

In general, ordinal utilities are not enough to explain the optimality of an action. Example:

	t_1	t_2
a_1	1500 €	200 €
a_2	0 €	300 €

$p(t_1) = .05$
$p(t_2) = .95$

Suppose that in this decision problem, the agent prefers more money to less money in proportion of the difference. Yet the following two ordinal utilities are equivalent representations of her preferences:

(4) $u_o(1500\text{ €}) = 1500,\ u_o(300\text{ €}) = 300,\ u_o(200\text{ €}) = 200,\ u_o(0\text{ €}) = 0$

(5) $u'_o(1500\text{ €}) = 400,\ u'_o(300\text{ €}) = 3,\ u'_o(200\text{ €}) = 2,\ u'_o(0\text{ €}) = 0$

One could be tempted to consider a_1 as the optimal action since it has the highest utility value. This, however, would ignore the agent's subjective probability p according to which she deems state t_2 much more likely to obtain than t_1. But in t_2, a_2 yields a better outcome. So, the optimality of an action should depend both on a utility and a probability. It should maximize an agent's expectation of its current utility, given her preferences and her subjective probability. "Optimality" in this sense is an expression of an action's expected utility.

At this point, it becomes clear why ordinal utilities won't do. Suppose we use them to construct the expected utilities as follows (the official definition follows below): For each state, we multiply the utility value of an action's outcome by the state probability (p) and then we sum the results, as in (6) and (7):

to von Neumann/Morgenstern is proposed. It is inspired by the work of Ramsey and requires less assumptions.

(6) “Expected utility” according to u_o

a_1: $u_o(1500\text{ €}) \times .05 + u_o(200\text{ €}) \times .95 = 1500 \times .05 + 200 \times .95 = 75 + 190 = 265$

a_2: $u_o(0\text{ €}) \times .05 + u_o(300\text{ €}) \times .95 = 0 \times .05 + 300 \times .95 = 0 + 285 = 285$

(7) “Expected utility” according to u'_o

a_1: $u'_o(1500\text{ €}) \times .05 + u'_o(200\text{ €}) \times .95 = 400 \times .05 + 3 \times .95 = 20 + 1.8 = 21.8$

a_2: $u'_o(0\text{ €}) \times .05 + u'_o(300\text{ €}) \times .95 = 0 \times .05 + 2 \times .95 = 0 + 2.65 = 2.65$

The problem is that u_o and u'_o are ordinally equivalent but according to u_o and p, a_2 is optimal while according to u'_o and p, a_1 is optimal.[11] This consequence is not acceptable. But we can use cardinal utilities to calculate the expected utility of an action given the probability the agent ascribes to the relevant states.[12] (The problem of the example is now solved as follows: Suppose that u_o and u'_o are cardinal utilities. (i) they are not cardinally equivalent since there is no positive linear transformation of one into the other. (ii) Of u_o and u'_o, only the former represents the agent’s preferences adequately; the utility values of u'_o do not reflect the comparative strength of her preferences.)

The example is a *decision problem under uncertainty.* In general, there is a finite set of actions A, a finite set of states T the world can be in, a probability p over T, and a set of outcomes O. To each action $a \in A$ in a state $t \in T$, an outcome $a(t) \in O$ is associated. We assume that the actions in A are feasible for the decision-making agent in question, that she has a cardinal utility u over the outcomes O and that p represents how likely she deems the states of T. Then the expected utility of an action a, denoted as $EU(a)$, is defined as the sum of the utilities for the outcomes of a in the different states weighted by the probability of the state, that is, $EU(a) = \sum_{t \in T} p(t) \times u(a(t))$.

The instrumental principle The way we’ve fixed the expected utility of an action suggests how the agent should answer her desires. If the value she ascribes to action a_2 is 285 but that for a_1 is only 265, then she ought to choose a_2. That is, an agent ought to choose an action whose expected utility is maximal. This decision principle is called the “instrumental prin-

[11] This is so since ordinal utilities are not additive.

[12] This is so since cardinal utilities and probabilities are additive. A so-called “Dutch book” argument shows that they should have this property (Jeffrey 2004:4–9).

ciple." It is the conjunction of SRB (for "subjectively rational behavior") and NDP (for "normative decision principle"):

SRB. An action a in a decision problem under uncertainty is *subjectively rational* for its agent iff a is among the actions with the maximal expected utility, that is, $a \in \{a' | \neg\exists a'' : EU(a'') > EU(a')\}$.

NDP. An agent in a decision problem under uncertainty ought to perform a subjectively rational strategy.

The normative decision principle highlights that we're dealing here with *prudential oughts.* This is of importance for Lewis' analysis of conventions. For these are the only *oughts* entailed by his proposal.

4.2.2 Strategic games

In the simple setting we'll use, a (strategic) game consists of (i) a set of agents (usually called "players"), (ii) for each player, a set of strategies, and (iii) a payoff function over strategy profiles. It is assumed that every player acts in a certain situation according to one of the strategies available to her. A *strategy profile* assigns to each player one of her strategies. *Payoff functions* are a special kind of utilities that satisfy the expected utility axiom.[13] Cardinal utilities satisfy this condition.

The BoS-example (3) is a strategic game. The players are Judith and Marc. Each of them has two strategies: Judith can go to Marc's friends ("U") or go to her friends ("D"). Marc can go to her friends ("L") or to his friends ("R"). Payoff functions that respect what I've said about their preferences are:

(8) $u_J(\langle U, L\rangle) = 2, u_J(\langle D, R\rangle) = 1, u_J(\langle U, R\rangle) = u_J(\langle D, L\rangle) = 0$

(9) $u_M(\langle D, R\rangle) = 2, u_M(\langle U, L\rangle) = 1, u_M(\langle U, R\rangle) = u_M(\langle D, L\rangle) = 0$

These claims determine a game since we have a set of agents, their strategies and their payoff functions. We can depict the game as follows (a so-called "normal form"):

[13] The expected utility axiom states that the utility of a gamble between two outcomes o_1 and o_2 with the objective probabilities p and $1 - p$ equals the sum of the utility of o_1 weighted by p and the utility of o_2 weighted by $1 - p$.

	L	R
U	$2,1$	$0,0$
D	$0,0$	$1,2$

The figure is to be read as follows. Judith is the "row-player." Her actions U and D appear as the headings of the rows. Marc is the "column-player." His actions L and R appear as the headings of the columns. Each cell of the table body represents a strategy profile which is determined by taking the headings of its row and its column as Judith's and Marc's action, respectively. The tuple of numbers of each cell denotes the respective payoffs. For example, the entry "$2, 1$" in the top left cell corresponds to the strategy profile $\langle U, L\rangle$ and means that Judith's payoff is 2 while Marc's is 1.

In a game, the individual decision problems (typically) depend on each other. Hence, optimality is ascribed to *strategy profiles* and not just to *individual strategies.* A central notion of optimality is the *Nash equilibrium.* According to it, a strategy profile is considered to be optimal in a game if no player can individually bring about a better outcome by doing her part of another strategy profile. Formally, this amounts to the following:

NE. A strategy profile s is a Nash equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) \geq u_2(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) \geq u_1(s'')$.

In a Nash equilibrium, no one can improve her situation by individually deviating from the action performed in the strategy profile, or equivalently, no one has an incentive to perform another action. A related equilibrium notion is the *strict Nash equilibrium* in which the $\geq$-signs are exchanged for $>$-signs:

SNE. A strategy profile s is a strict Nash equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) > u_2(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) > u_1(s'')$.

In a strict Nash equilibrium, each agent would do worse if she were to deviate. Obviously, every strict Nash equilibrium is also a Nash equilibrium.

The BoS-example has two strict Nash equilibria: $\langle U, L\rangle$ and $\langle D, R\rangle$. Judith prefers the first to the second and *vice versa* for Marc. Nevertheless, both strategy profiles are strict Nash equilibria.

If a game has more than one Nash equilibrium, then this has consequences for coordination. Doing one's part in a Nash equilibrium is a necessary condition to bring about an optimal outcome.[14] But it is not sufficient: the strategy profile $\langle U, R \rangle$ satisfies the condition but is not optimal.

This illustrates an important point about classical game theory: it has little to offer in games with multiple equilibria since the equilibrium notions do not single out exactly one equilibrium.[15] An account of conventions can be understood as a contribution to this problem. It's not a complete solution since conventions exist only in situations in which *common interest* dominates.

4.2.3 Coordination games

An obvious proposal to explicate what it is to be *in the common interest* is this: strategy profiles that are SNEs are in the common interest. For in an SNE, no one wants to deviate. Lewis demanded more: a strategy profile has to be an SNE *in which nobody wants anyone else to act differently while oneself acts as one does*:[16]

PCE. A strategy profile s is a proper coordination equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) > u_2(s')$ $\underline{\text{and}}$ $u_1(s) > u_1(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) > u_1(s'')$ $\underline{\text{and}}$ $u_2(s) > u_2(s'')$.

In other words, a proper coordination equilibrium (PCE) is a strategy profile in which everyone wants everyone to act as they act. No one wants that someone else deviates. A good example to illustrate the difference between PCEs and SNEs is the *game of chicken* (figure 4.1(b)). One story for that game goes like this. Two guys want to prove themselves and do so by driving their cars towards a cliff. The one who stops first is then the "chicken" because he has chickened out ("C"). The one who speeds longer ("S") wins. In this game, it is for each driver better if the other chickens out first, but both prefer that both chicken out to both speeding longer. For otherwise,

[14] There is another sense of "optimal" in which it might not be necessary. The Prisoner's Dilemma illustrates this; see (Kuhn 2007). For this reason, one should probably rather speak of "stability" to characterize the role of NEs.

[15] This problem is known as the *equilibrium selection problem* (Alexander 2009:§3.1).

[16] *Cf.* (Lewis 2002:8–24). He gives a more general definition for n-player games.

both cars would fall down the cliff.[17]

	L	R
U	2, 1	0, 0
D	0, 0	1, 2

(a) The BoS-game

	C	S
C	8, 8	2, 9
S	9, 2	0, 0

(b) The game of chicken

	A	B
A	2, 2	0, 0
B	0, 0	1, 1

(c) A coordination game

Figure 4.1: Contrasting PCEs (a) with SNEs (b) and with (c) Pareto-optimality.

The game of chicken has two SNEs, namely $\langle C, S\rangle$ and $\langle S, C\rangle$. But in every equilibrium, one of the agents wants the other to change her action. For example, if I am about to chicken out and you're about to speed longer, then we end in the equilibrium $\langle C, S\rangle$. Nevertheless, I still want you to chicken out since I prefer the strategy profile $\langle C, C\rangle$ to $\langle C, S\rangle$. Hence in such a game there is no PCE. Clearly, in chicken-like situations, there is a strategic incentive to deceive and to deviate in the SNEs.

Observe that a PCE is not the same as a Pareto-optimal NE.[18] In the game 4.1(c), both strategy profiles $\langle A, A\rangle$ and $\langle B, B\rangle$ are PCEs but only the former is Pareto-optimal.

From PCEs it's only a short step to Lewis' formalization of a coordination problem. If a strategic game has at least two PCEs, then there is common interest and a need to coordinate. Let us define a "coordination game" as a strategic game having two or more PCEs.[19]

Lewis is not very explicit about his motivation to define coordination games in terms of PCEs. The best Lewis offers is in *Convention* where he says that he wants "to confine [his] attention to situations in which coincidence of interest predominates" (Lewis 2002:14). From this it does not follow that SNEs wouldn't do. I think they do and count the SNEs in chicken as ways to bring about coordination. Here, Lewis and I disagree

[17] This and other popular versions of this game can be found in (Wikipedia 11.03.2009).

[18] Among the NEs in a game, an NE $\langle s_1, s_2\rangle$ is *Pareto-optimal* iff there is no other NE $\langle s_1', s_2'\rangle$ in the game such that either $(s_1' \geq s_1'$ and $s_2' > s_2')$ or $(s_1' > s_1'$ and $s_2' \geq s_2')$.

[19] PCEs in coordination games relate as follows to his official definition of a convention: If we reconstruct a normal form game from the description of R, its alternatives, and the constraints imposed on the members' preference structure in clauses 3, 4, and 5 of the official definition, then the game is a coordination game in which R corresponds to a strategy profile which is a PCE.

about *what it is to be in the common interest* or, in other words *what it is to coordinate.*[20] The consequence of my proposal is that I'm willing to say that in the chicken game there can be conventions. One could be that the younger driver always chickens out first. As long as one gains something by playing the game compared to not playing it, it's still individually beneficial to enter the game.

4.2.4 Conventions in game theory

Suppose that there are recurrent situations realizing a coordination game. In these situations, the players always conform to a particular PCE. It seems apt to call this regularity a "convention." But the characterization is not general enough as the car driving example (2) illustrates: While typically two drivers coordinate in a realization of a game, it is not always the same two who coordinate; they are drawn from a bigger *population.*[21]

To accommodate such cases, a subtle reinterpretation of the game theoretic formalism is required: The *set of agents* of a game is now understood as the *set of roles* agents can have if they play the game.[22] So, for every play of a game we need now an assignment of roles to the agents playing it.[23]

Now we can define with Lewis the game theoretic notion of a convention:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a convention if and only if, in any instance of S among members of P,
>
> 1. everyone conforms to R;

[20] Some economists go further and do not only count SNEs in games with at least two SNEs as ways to bring about coordination. Robert Sugden (1998:381) reports that economist are willing to say that any NE (that perhaps satisfies some further stability condition) in a game with more than one NE can be a convention (and hence is a way to bring about coordination). This fits evolutionary accounts that use the notion of an evolutionary stable strategy; see §7.2.

[21] Lewis implicitly assumes that the members of a population don't change over time. This is restrictive. But for simplicity I follow him.

[22] The point is obvious but I know of no text on game theory where it is made explicit.

[23] The assignment is restricted with regard to two conditions: (i) The players need to be equipped with strategies enabling them to play the role and have strategies of the form "If I am in role of the row player R, then I do .... But if I am in role of the column player C, then I do ...". (ii) An agent can only play a role if her preferences are equivalent to those associated to the role in the game.

2. everyone expects everyone else to conform to R;
3. everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a proper coordination equilibrium in S. (Lewis 2002:42)

This definition seems simpler than the ones Lewis proposed (§4.1). Yet Lewis gave it up since it rules out cases which should count as a convention. His example is (Lewis 2002:46 ff.):

(10) A product is produced by two companies. Both companies have a common interest in profitable prices, yet no one wants to be much more expensive than the other. An agreement to sell at the highest price would be optimal for them. But by law, they may not do so. So, they have to adopt an informal scheme of fixing prices, one of them being: "Follow the company fixing the price first and set a slightly lower price." Such an informal scheme is a convention between the two companies.

In this scenario, setting prices is a *continuous activity* possibly requiring slightly different price setting actions in different situations. According to the current proposal, we have to divide the life time of the convention into a series of *independent* realizations of a coordination game. But is that possible? If we choose a small period of time for each realization, say a business day where both companies set the prices in the morning, then we miss an important part of the coordination, namely the consumer reaction to the pricing decision which is a determinant for the companies' payoffs.

Using a larger time period also doesn't work. If we use *actions* as the strategies the companies can choose, then they act more than once within such a time period. But this is in conflict with the idea that in a strategic game, each player acts only once.

If we use *strategies* chosen by the companies in the beginning once and forever, then we do not do justice to the fact that on each price setting occasion, the companies have to coordinate.

Lewis' moral is that the game theoretic definition is not adequate. Hence Lewis (2002:68–71) gave it up and restated his analysis in the way presented earlier in §4.1. The crucial move is from (11) to (12):[24]

[24] There are some further changes leading to the definitions Lewis endorsed; most importantly *common knowledge* which I discuss below in §4.3.1.

(11) Each reoccurring situation of type S must be a self-contained coordination problem which satisfies the clauses 1 to 3 above.

(12) The situations together must form a coordination problem; the individual situations needn't be a self-contained coordination problem.

The move comes at a cost, even if it's not as drastic as Guldborg Hansen (2009:109) wrote: It "gives up all the intuitive, technical and theoretical understanding" of game theory. Contra Lewis, it is not required if we follow Syverson (2003:§10.3). He changes (11) to (13):

(13) Each reoccurring situation of type S is a subsituation of type S' which satisfies the clauses 1 to 3 (in clause 2 and 3 "S" has to be replaced by "S'").

The subsituations S in Lewis' price-setting example are situations in which one company fixes the price.[25] Situations of type S are not self-contained coordination problems. But S is a subsituation of type S', which is plausibly a type of situation which includes the other's company fixing the price. Hence, situations of type S' are coordination problems. So, there is still a realization of a coordination problem in each reoccurring situation.[26]

Consequently, Lewis' game theoretical definition is more viable than he thought. For parity with his official definition, the regularity in *behavior* can be extended to regularities in action *and attitudes* and we could allow for some exceptions. The result is that there are two viable definitions of a convention – the official definition we considered first and the game theoretical definition with Syverson's modification.

4.3 Evaluation

Lewis' detailed analyses has received a lot of criticism. Before we evaluate Lewis' account, I'd like to take stock.

A quick check would show that the three definitions of a convention we've discussed satisfy the pre-theoretic characterization of rational and

[25] Syverson's proposal makes use of situation semantics (Barwise and Perry 1986) but arguably, one could restate it in terms of events.

[26] A consequence of Syverson's proposal is that it depends now also on the choice of S' whether a regularity is convention or not. If S' is too inclusive, then there might be no regularity anymore. But this seems to be as it should.

rationally justifiable conventions (§1.2), on the respective interpretations of agents as either deliberating or non-deliberating rational agents. On exegetical grounds, I think that both interpretations can be attributed to Lewis.

Dispositional conventions, however, do not fit the picture. For at least the following condition must be satisfied: If an agent performs an action that is elicited by a non-deliberative mechanism, then the agent could have deliberated about what to do and if she did, then she would have come to perform the same action. Agents not capable of deliberating don't satisfy the condition, even if they have behavioral dispositions to act convention-conformingly.

Lewis only endorsed what I've called the "official" definition of *Languages and language* (Lewis 1975) (§4.1). It turned out that the game theoretical definition is also defensible, using the suggestion of Syverson to restate it in terms of situations that are a subsituation of another one (§4.2.4). So, there are two Lewisian contenders.

In this section, I start by discussing four topics requiring changes to Lewis' proposal: common knowledge (§4.3.1), alternatives (§4.3.2), regularities (§4.3.3), and the explanation of convention-conforming behavior (§4.3.4). On that understanding of Lewis' account, I draw an interim summary in 4.3.5 and then discuss some objections from the debate about Lewis' account.

4.3.1 Common knowledge

Lewis thought that the clauses 1 to 3 of his game theoretical definition were not sufficient and added a further condition: everything about the convention is common knowledge among the members of the population in question.

From the perspective that parties to a convention are deliberating agents, clause 2 (each believes that the others conform) is understandable: To act optimally, an agent needs to be well informed. But given that each party to a convention (truly!) believes that the others conform to it, why do we need common knowledge? What is common knowledge in the first place? I start with the second question. On that basis, I discuss the first.

Lewis is usually considered to have been the first person explicitly defining common knowledge. But what he defined as "common knowledge" in his book *Convention* (Lewis 2002:52–68) is not what people in philosophy, game theory, computer science, ... usually take it to be. The *basic idea* of the usual proposals[27] is that there is common knowledge among a group G of agents that p iff for all agents A, A' of G:

- A knows that p, and
- A knows that A' knows that p, and
- A knows that A' knows that A knows that p, and
- so on, *ad infinitum*.

There are different definitions based on this idea but all agree that common knowledge is *factive* since it is defined in terms of factive knowledge. That is, if among G it is common knowledge that p, then p. But contrary to what many believe,[28] common knowledge in Lewis' sense ("Lewis common knowledge") is not factive. It is possible that in a group there is Lewis common knowledge that p, and yet p is false.[29] The point is important for two reasons. First, criticism against the usual notion of common knowledge does not necessarily carry over to Lewis common knowledge. Second, the roles the usual notion can have (in explanations, justifications, ...) cannot necessarily be played by Lewis' notion.

But if Lewis common knowledge is not knowledge, what is it then? He defined it as follows:

> [I]t is common knowledge in a population P that ___ if and only if some state of affairs A holds such that
>
> 1. Everyone in P has reason to believe that A holds;
> 2. A indicates to everyone in P that everyone in P has reason to believe that A holds;
> 3. A indicates to everyone in P that ___. (Lewis 2002:56)

Example: Between you and me (P) it is Lewis common knowledge *that we drive right* (___). Then by definition a state of affairs A obtained such

[27] See (Vanderschraaf and Sillari 2007) for a helpful overview. Aumann (1976) provided an important definition of common knowledge within epistemic logic. Gerbrandy (1999:45) compares several important notions and proves their equivalence under reasonable finiteness assumptions. In the recent literature, Sillari (2008) is one of the few who distinguishes Lewis' definition from these other proposals. For example Montet and Serra (2003:5) think that Lewis' definition is the same as the usual one and so does Binmore (2008:22).

[28] Typical examples are (The Royal Swedish Academy of Sciences 2005), and (Vanderschraaf and Sillari 2007), and (Wikipedia 28.01.2009).

[29] Andreas Kemmerling is, to my knowledge, the only person who has ever made the point that Lewis common knowledge is not a kind of knowledge. See (Kemmerling 1976:24).

that conditions 1 to 3 are satisfied. A is the state of affairs consisting in us crossing each other on a street. Thereby, both of us have reason to believe that we cross each other on a street (condition 1). Our crossing on a street indicates to us that we have reason to believe that we cross each other on a street (condition 2). Moreover, our crossing indicates to us that we drive right (condition 3).

The example shows that "knowledge" is nowhere used, only *reasons to believe* and *indications*. So, if Lewis common knowledge were factive, then it would be a consequence of these notions.

It does not follow from me having a reason to believe that p that I actually believe that p. This only follows, if we additionally assume that I am rational (in the CK-supporting sense explained below). To have a reason to believe that p is roughly to be justified to believe that p.[30]

Even if we assume common rationality (that is, everyone is rational in the required sense and assumes that everyone is rational in this sense and so on ...), all we can derive is that agents in P believe that we drive right. But beliefs are not factive. Neither are indications. Indications provide a justification-transfer mechanism and are defined as follows:

> [l]et us say that A indicates to someone x that ___ if and only if, if x had reason to believe that A held, x would thereby have reason to believe that ___. (Lewis 2002:52 ff.)

Without going into the details of this notion, observe that the only epistemic (or justificatory) notion is *having a reason to believe*. Consequently, we can't derive knowledge from indications either. Hence, Lewis common knowledge is not factive and thus is different from common knowledge according to the basic idea. A better name would have been "common reason to believe."[31]

[30] A different interpretation has been proposed by Cubitt and Sugden (2003:184) who provide an axiomatic reconstruction of Lewis common knowledge in terms of knowledge. They think that one has reason to believe that p iff one accepts, "as a normative standard," some logic of reasoning such that – in the logic – p is either treated as self-evident or derivable from propositions that are treated as self-evident. – I guess they want to say that one has a reason to believe that p iff one accepts a norm according to which one ought to believe that p. I think they are wrong about Lewis. For one, Lewis does not mention "normative standards" or "self-evident" propositions in the relevant context.

[31] Lewis (1978:44 fn. 13) was aware that calling his notion "common knowledge" is misleading. In *Languages and language* (Lewis 1975:6), after *Convention*, Lewis had the goal to weaken the requirement by allowing also "merely potential" or "negative" knowledge (the

To derive that some proposition p is common knowledge in a population P from Lewis common knowledge, p has to be true, and the members of P have to satisfy the following conditions: (i) Each is rational in the sense that if she has a reason to believe that p, then she believes that p. (ii) Each believes that the other players are rational in this sense. (iii) They have common inductive standards and (iv) a common background information.[32] These "CK-supporting" conditions are arguably *idealized*, since for complex propositions it is implausible that agents can satisfy (i) and (ii); also (iii) and (iv) are hardly satisfied in bigger groups.

Generality *in sensu composito* and *in sensu diviso* Lewis tried to weaken the conditions for common knowledge. To this end, he invoked a distinction between two kinds of generality, namely generality *in sensu composito* and generality *in sensu diviso*.[33] Consider the following example: "Whenever I drive in Switzerland, I want to drive right." The two kinds of generality are now this:

- *In sensu composito*, I have one want with general content, namely the want that whenever I drive in Switzerland, I drive right.
- *In sensu diviso*, I have many wants with nongeneral contents, namely the want of *this* particular situation that I drive right, the want of *that* particular situation that I drive right, and so on and so forth.

Attitudes of the two kinds are compatible; an agent can have both of them. They differ in their entailments. *In sensu composito* the want in the example is consistent with the proposition that when I drive in Switzerland, I don't want to drive right. I might fail to recognize that the situation in which I am is one of driving in Switzerland. And even if I recognize the situation as such, I might fail to infer from the want with the general content the particular want with nongeneral content, namely that I want to drive right in that case. Contrast this with generality *in sensu diviso*. Here, whenever I am in a relevant situation, I want to drive on the right, whether I recognize it as being relevant or not. Claiming the opposite would be inconsistent.

requirement not to have certain knowledge). But it is also stronger since he characterizes common knowledge as a kind of *knowledge*. I think that Lewis was maneuvering.

[32] See (Lewis 2002:51–57).

[33] See (Lewis 2002:64–68). The distinction is intimately related to the *de dicto*/*de re* distinction.

In which sense should we understand common knowledge? Lewis' answer is "whichever kind it is that ensures the agent's ability to apply his general attitudes to the instance at hand. And that is a limited generality *in sensu diviso*" (Lewis 2002:66). Lewis points out that what we want to claim is that in every particular situation, the agent in question has a certain attitude. General attitudes *in sensu diviso* ensure that; attitudes *in sensu composito* don't – for the reasons mentioned above (due to failures to recognize and to infer). So, we only need to assume that they have general attitudes *in sensu diviso*. This is Lewis' position with respect to all attitudes involved in the analysis of conventions.

Roles of common knowledge In Lewis' analysis of conventions, the notion has the following jobs (I return to them below):[34]

First, Lewis thinks that it is a common feature of his central examples. For descriptive adequacy it should be included in the definition.

Second, common knowledge plays a justificatory role (§4.3.4). Common knowledge entails a hierarchy of reasons to believe. Higher-order beliefs justify the lower-order ones.

Third, Lewis wanted to rule out cases where people are wrongly motivated by having *false beliefs*. A particular case Lewis (2002:59) wants to rule out is one where people drive right because they falsely believe *that all the others except for themselves drive right habitually* and the best response to that is also to drive right. In this case, they do not "really" coordinate. The common-knowledge assumption, together with some rationality, rules it out: By common knowledge there is a state of affairs indicating that their belief is false. By rationality, they believe that what they believed is false.[35] Hence, they no longer have a good and decisive reason to drive right. Thereby, common knowledge plays the role of a *transparency condition*.

Fourth, common knowledge ensures stability since it creates reasons to do one's part of the conventional regularity and thereby reinforces conform-

[34] *Cf.* (Schwarz 2009:191 fn. 2). I ignore a further role: to rule out cases which have only especially bad alternatives which just satisfy the condition of being generally conditionally preferred (condition 3 and 4 in the final definition below). Lewis argues that such cases do not count as conventions under his analysis since the alternatives are not common knowledge (Lewis 2002:73 ff.). I don't share the intuition. An alternative is an alternative, after all.

[35] In general, Lewis common knowledge is not factive. In this case, Lewis is arguably right. For it would be irrational to uphold the false belief. Hence, rational agents revise it.

ing behavior (Lewis 1975:6).

Fifth, the common-knowledge assumption is used to show that his conventionalist account is Gricean, that is, that it entails that speakers normally speaker-mean something (see §5.3.4).

Criticism There are several issues with the common-knowledge assumption. First, common knowledge is a fragile concept. To *preclude* that some proposition p is common knowledge in a group, it suffices that a single member has doubts whether all members of the group satisfy the conditions for common knowledge that p. For to have doubts that p is not to know that p. But for p to be common knowledge, everyone has to know that each member knows that p. (I think appealing to the *ceteris-paribus* that only almost all have to have common knowledge is not coherent with the notion of commmon knowledge.)

The argument has to be modified for Lewis common knowledge. Plausibly, in a situation where someone has such a doubt, there is no state of affairs A indicating to everyone in the group that each member has reason to belief that A obtained. But then p cannot be Lewis common knowledge since condition 2 is violated according to which A indicates to everyone in P that everyone in P has reason to believe that A obtained. So, it's doubtful that in bigger groups the common-knowledge assumption can be true.[36]

The consequence is that in many cases we'd like to count as a convention, the common-knowledge assumption is plausibly not true. Hence, it cannot be a necessary condition for there to be a convention.

Second, Burge (1975:250) and Kemmerling (1976:117–123) independently argued that a community can rightly be said to conventionally use a language even if they are not aware of the possibility that there are other languages. So, it seems that there can be conventions without the alternative being known (this is required by the common-knowledge assumption and condition 5 of the definitions in §4.1).

Third, as Burge (1975:250 ff.) pointed out, on Lewis' account, members of a convention cannot find out that they are members of a convention: if they are, then they commonly believe that (under weak rationality assumptions). This is counter-intuitive since members of a convention can *discover* that there is a convention among them.

[36] Binmore (2008:23) argues similarly.

The three objections converge on issues with common knowledge. Luckily, the common-knowledge assumption can be given up. I return now to the roles:

First, as Savigny (1985:87 ff.) observed, common knowledge is not a common feature of ordinary conventions; often when conforming to a convention, we behave habitually without knowing why. Speaking one's first language is a case in point.

Second, why should we *require* that the agents' beliefs are justified? Consider again the rowing-a-boat case (1): As long as you and me want to glide straight and achieve that by coordinating our behavior in a certain way, there is a convention between us, *whether our beliefs are justified or not*. Moreover, based on the agents' desires, their actions are justified.

Third and for similar reasons, it's not obvious whether one really wants to rule out Lewis' car-driver scenario.

Fourth, common knowledge is not required for stability. There can be stable regularities without there being common knowledge. This is one of the lessons learned from evolutionary game theory. Skyrms (1998) showed convincingly that there can be stable signaling regularities among agents having no common knowledge.

Fifth and finally, it's not obvious that being Gricean is a desirable feature. I'm going to argue that it is not (chapter 6). Hence, common knowledge is not required to have this role (and as I argue in §5.3.2, the conventionalist theory can be made Gricean more directly by stating the communicative regularities directly in terms of speaker-meaning).

So, it seems that we can do without common knowledge. Can we do without any epistemic conditions at all? There is still condition 2 requiring that parties to a convention believe that the others conform. This seems to me a plausible condition for rationalistic and rationally justifiable conventions. For in such cases, we want the parties to be epistemically aware of the coordination problem they solve. In case of dispositional conventions, I think one can drop condition 2 as well. Consequently, in such cases, the agents needn't conceptualize the situation as a coordination problem. Millikan's account of conventions (chapter 7) is of this kind.

4.3.2 Alternatives

Conventions are arbitrary in the sense that there is at least one non-trivial alternative R' to a conventional pattern R that could have prevailed as well because the belief that the others conformed to R' would have given parties to the convention a good and decisive reason to conform to R'.

But Lewis' characterization of alternatives in terms of good and decisive reasons is unclear. Tyler Burge's discussion of the sentimental-hat-tippers case (Burge 1975:251 ff.) invites a clarification: a group of sentimental hat tippers "would rather fight for the traditional greeting [...] than switch to another [way of greeting strangers]" (p. 252). According to Burge the case should be counted as a convention, even if it does not clearly satisfy Lewis' conditions.

A first interpretation of "giving good and decisive reasons" Burge suggests is in terms of *motivational efficacy*. Since the hat tippers are not motivated to switch, the case is not a Lewis convention; hence, so Burge, we shouldn't use this interpretation.

The second interpretation is in terms of *rational sufficiency*. But it is unclear whether the sentimental hat tippers would be *irrational* in preferring not to switch to another greeting practice if the others did. Hence, it could be that such a reason is not rationally sufficient to conform to the alternative R'. To correct this, Burge considers conditioning the alternative-clause on "willingness to continue to participate in a communal practice that serves substantially the same social functions as the original one" (p. 253). One problem of this condition is that it presupposes that the agents could switch to conforming to the alternative – in some cases this is too demanding (p. 254). For example, if using a language is conventional, then plausibly many users couldn't switch to another one simply because they are not good learners (*cf.* p. 250).

I think Burge's points are well taken. They highlight an unclear point of Lewis' analysis. A more plausible proposal about conventional regularities and their alternatives which does not have the indicated problems is this: conventional regularities are learnable (and thereby feasible) and "historically accidental." The condition that there is at least one incompatible alternative should be so understood that the alternative could have been learned and served the same goal as well (*cf.* p. 254). I think that this is a welcome clarification of Lewis' proposal.[37]

[37] This is contra Peacocke (1976:169 ff.) who thinks that one should drop the alternatives-clause. I prefer to change the meaning of "alternative."

4.3.3 Regularities

According to Lewis, conventions are a kind of *regularity.* But what is regularity and how should we interpret the talk of "regularity"? To my knowledge, there are no answers in Lewis' written work. At some point, Margaret Gilbert also wondered and asked Lewis. She reports that Lewis takes regularities to be *pairs of properties* (Gilbert 2008:8) of the form $\langle F, G \rangle$. Plausibly, the properties are properties of events and hence, the proposal can be understood as "all F-events are G-events."[38] I'll argue that this proposal is too strong; for we can allow for a substantial share of Fs that are not Gs. But first, I'd like to consider the question how one could understand "regularity". This relates to desideratum DesC2 for an adequate account of conventions according to which this question should be answered (§1.4).

Hence, let us consider the relevant senses of the word "regularity" and find out whether there is one which suits Lewis' needs. There seem to be three relevant senses of the word "regularity": (i) *rules*, (ii) *conditional probabilities*, and (iii) *patterns of activity.* Textual evidence suggests that Lewis does not want to endorse the first option since he wanted to contrast conventions, a kind of regularity, with rules.[39] So, options 2 and 3 remain.

I take the received view to be that regularities are high conditional probabilities. For example, this assumption is implicit in the objections of Millikan (1998) and Kölbel (1998) against Lewis' account. I'll argue against this view and propose to understand them as patterns of activity. For understanding Lewis' "regularity" in the sense of a high conditional probability has certain strange consequences. To prepare the argument, I introduce the notions of a high conditional probability and of a pattern of activity. Then the argument is provided. Finally, I discuss an objection by Millikan that the degree of conformity can be much lower than Lewis suggests.

Regularities as high conditional probabilities In this sense, regularities are high conditional probabilities between event types F and G. We can

[38] Hence, "regularity" means here the same as in regularity theories of laws of nature.

[39] In *Convention*, Lewis has a chapter called "Convention contrasted" and section 4 of the chapter has the title "Rules." In this section, Lewis distinguishes rules from conventions, see (Lewis 2002:100 ff.). Moreover, in a later publication called *Meaning without use* (Lewis 1992:109 ff.), Lewis indicates that he does not want to use the notion of a rule in the context of conventions.

express this more precisely as follows: the objective conditional probability of F given G is close to 1, *i.e.* $P(F\,|G) \approx 1$ and $P(F\,|G)$ is substantially bigger than $P(F)$. In the BoS-example (3) about Judith and Marc's lunch coordination problem, possible conditional probabilities are the following: (i) Whenever Judith and Marc want to have lunch together (G), both go to her friends (F); (ii) Whenever Judith and Marc want to have lunch together (G), they go to his friends, respectively (F').

Regularities as patterns of activity The notion of a pattern of activity I'm interested in here is such that the thunder after the lightning is *not* a pattern of activity but *activities* of agents are. Patterns of activity are a type of activity. The basic idea is that we abstract over certain features of a class of events to derive the type of activity: the particular agents involved (but we keep their roles in the event), the times when certain sub-events happened (but we keep their temporal relations), and so on. Hence, a pattern of activity is a type of activity which normally involves two or more agents in certain roles. An *activity* is an event consisting in a sequence of possibly overlapping doings of at least one agent and can also include other states of affairs such as mental states (*e.g.* attitudes), their transitions, and wordly events not involving agents.[40] This way of understanding regularities might be technical but I think we're all quite competent in using the expression "pattern", *e.g.* when someone says that there is pattern in the ways people cross streets, approach each other to make out, and so on.

Against high conditional probabilities First, in all definitions of "convention" Lewis offers, conventions are a kind of regularity. The definitions' first clause is something like "Everyone conforms to regularity R." But if there is a high conditional probability of R-conforming behavior given some type of situation, why do we need this clause at all?[41] The first clause makes sense if we understand "regularity" as a *pattern of activity.* Then (i) conventions are a kind of a pattern of activity R and (ii) R is conformed to by everyone. The high conditional probability would be a consequence (in cases where non-conformity is not widespread).

[40] Millikan conceives of conventions as a kind of such patterns; see chapter 7.

[41] Davidson (2001:276) was to my knowledge the first who observed this quirk.

Second, it should make sense to talk about a regularity R whether it is exhibited or not, that is, whether the conditional probability is around 1 or much lower. The existence of a convention implies that there is an alternative regularity R' which is not actually conformed to. Hence, there is no high conditional probability with respect to R'. But then, it makes no sense to say that R' is a regularity in the *conditional-probabilities* sense. Understanding "regularity" as a *pattern* works again. Moreover, if a conventional "regularity" is not so strictly conformed to, then there is also no high conditional probability. Still, there is something members sometimes conform to and deviate from. That thing is a pattern of activity.

Consequently, my proposal is to understand the talk of "regularity" in Lewis' definitions of "convention" in the sense of a *pattern of activity.*

Degree of conformity Understanding Lewis' regularities in the sense of high conditional probabilities settled the question to which degree the parties to a Lewis convention conform to it. On the pattern-of-activity interpretation, it becomes a question to be answered. Arguably, the received view is that the parties almost always conform to a Lewis convention. Ruth Millikan has objected. Conventions need not be regularly conformed to:

> That people need not regularly conform to noncoordinating conventions is clear. Few actually hand out cigars at the birth of a boy, nor does everyone wear green on St. Patrick's Day, or decorate with red and green on Christmas, or punt from the deck when on the Cam River. (Millikan 1998:170)

While her examples are about *non-coordinating* conventions, the points carries over to "proper" coordination conventions (the ones Lewis studied). On Millikan's proposal (chapter 7), the parties to a convention only have to conform *often enough.* On Lewis' official definition, the degree of conformity has to be *almost perfect.*[42] According to the received view, Lewis cannot make a substantial concession with respect to the degree of conformity. For example, Kölbel (1998:304–308) explicitly claims so.

It seems to me that the common lore is wrong on that count. We can weaken Lewis' second definition along the following lines (incorporating the

[42] There is a version of his first definition that allows for conventions whose regularities are conformed to only in a fraction of cases (Lewis 2002:78 ff.). I think that this version was only meant to address the vagueness of the "almost"-qualifications and not to make room for widespread non-conformity.

suggested changes):[43]

A pattern of activity R, in action or in action and attitude, is a *convention* in a population P if and only if, within P, the following five conditions hold. (Or at least they almost hold. A few exceptions to the "everyone"s can be tolerated.)

1. Everyone conforms to R *often enough.*
2. Everyone believes that the others conform to R *often enough.*
3. This belief that the others conform to R *often enough* gives, *ceteris paribus*, everyone a good and decisive reason to conform to R himself.
4. *Ceteris paribus*, there is a general preference for general conformity to R rather than slightly-less-than-general-conformity.
5. R is not the only possible pattern meeting the last two conditions. There is at least one alternative R' which could have prevailed as well such that the belief that the others conformed to R' *often enough* would give, *ceteris paribus*, everyone a good and decisive reason to conform to R' likewise; such that, *ceteris paribus*, there is a general preference for general conformity to R' rather than slightly-less-than-general conformity to R'; and such that there is normally no way of conforming to R and R' both.

The two crucial changes are: (i) Clause 1 now only requires *often-enough* conformity. In terms of expected utilities it has a determinate content. Roughly, for every member m and alternative pattern R'': m's expected utility to conform to R must be higher (by some margin) than her expected utility to conform to R'' where R'' is like R except she or some other agent is behaving differently. (ii) Clauses 2 and 3 are hedged under a *ceteris-paribus* assumption. This allows that in a particular situation, the agents prefer not to deviate from R. I think that it's not possible to list the conditions in which this could be so. But we can say at least this: The agents' conditional preferences to conform must be so strong that clause 1 is satisfied.

The new proposal is compatible with agents always maximizing their expected utilities. What may vary are the agents' preferences in particular situations.

According to the proposal, the cigar case can be a convention if, amongst other things, the members of the respective population have, *ceteris paribus*, a general preference for handing out cigars at the birth of a boy. On particular situations of a boy's birth, the relevant parties are allowed to have the opposing preference not to hand out a cigar. That is to say that the

[43] The game theoretical definition can also be changed in a similar way; the basic idea is that in a coordination game, SNEs in mixed strategies can be conventional regularities.

new proposal tolerates a certain share of boy-birth events in which the respective agents' all-things-considered preferences are so as not to conform to the pattern of activity of the convention. How big this share is depends on the agents' expected utilities but it can be substantial.[44]

A consequence of accepting this modification is that expected utilities play a more important role. (For this reason I call it the *expected-utility* approach.) The moral then seems to be the following: Millikan's objection can be answered by modifying Lewis' definitions.

4.3.4 Explaining convention-conforming behavior

The discussion of common knowledge (§4.3.1) suggests that Lewis assumes that agents are rational. But he maneuvers between positions of two extremes:[45]

- Parties to a convention rationally deliberate. Their actions are brought about by a deliberative mechanism which elicits behaviors on the basis of their beliefs, desires, and the instrumental principle (§4.2.1).
- Parties to a convention do not deliberate. Their actions are brought about by a non-deliberative mechanism, *i.e.* behavioral dispositions such as habits.

According to the first position, parties reason about what to do. According to the second position, there are no such reasonings. This distinction is clear cut but arguably too clear cut. The topic is a delicate one since it bears upon the nature of beliefs and desires and their role in explanations of actions.[46]

Hence, I should be clear about the goal of my discussion, namely to highlight a certain tendency in Lewis' writing that is implausible.

Exegetical digression I think that there is no conclusive textual evidence to attribute to Lewis one of the two positions and plausibly, as we will see below, he holds a position that is in between. Yet, Savigny (1985:87) seems mostly right in interpreting Lewis as leaving "no room for doubt that

[44] Any degree of conformity above 50% is sufficient in symmetric pure coordination games; in asymmetric games, the degree can be even lower for some players.

[45] As far as I know, Kemmerling (1976:120–122) was the first to observe this.

[46] Wolfgang Schwarz pointed out to me in p.c. that for a Ramseyian (and Rylean) like Lewis beliefs and desires can be part of explanations of actions even if they are performed habitually without any (explicit or implicit) deliberation. According to such a conception, beliefs and desires *are* behavioral dispositions.

[...] all but 'children and the feeble-minded' are sufficiently rational" (in the deliberating-sense). Going by quotes, we find support for both positions. The following supports the first position. It occurs where Lewis explains how coordination games are solved.

> Agents confronted by a coordination problem may [...] succeed – if they do – through the agency of a system of suitably mutual expectations. Thus in example (1) I may go to a certain place because I expect you to go there, while you go there because you expect me to [...]. In general, each may do his part of one of the possible coordination equilibria because he expects the others to do theirs, thereby reaching that equilibrium. (Lewis 2002:24 ff.)

In the context of the quote, Lewis considers that agents come to act convention-conformingly on the basis of deliberating about what to do. This deliberation includes the replication of the other's deliberation about what to do. So, an agent reasons back and forth, thereby justifying her plan to act in a certain way.

Other passages suggest that agents need not be deliberating as long they act rationally and could deliberate. In the context of his Signaling Games theory Lewis writes:

> But we do not have to represent the agents' actual reasoning. We have to consider only the rational justifications of their choices by practical reasoning they *could* go through, given their beliefs and desires. Yet this is not to renounce an interest in explaining their choices. Justifications do explain choices, whether or not the agent actually goes through a process of reasoning following the justification. For it is a fact of human nature that we tend to act in ways justified by our beliefs and desires, even when we do not think through the justification. I may put it negatively: whatever may be the habitual processes that actually do control our choices, if they start tending to go against our beliefs and desires they soon would be overridden, corrected, and retrained by explicit practical reasoning. (Lewis 2002:141)

Consequences Depending on the position taken, Lewis' account (i) explains and justifies the agents' behavior in terms of rational deliberation (first position), (ii) explains their behavior in terms of non-deliberative mechanisms (however they are brought about) and justifies the behavior as subjectively rational behavior (second position), or (iii) is in between. Let us consider the two extremes (i) and (ii) to see why the third option is most plausible. (It's arguably the one Lewis endorsed, see footnote 46 on page 116.)

Lewis should not endorse the first position. For there are other (non-deliberating) mechanisms which can explain why agents act rationally. A plausible candidate is routinized behavior: recurring actions in human agents tend to become routinized. According to social cognition, a standard view in social psychology (Fiske and Taylor 1991), routinized actions are elicited by half-automatic processes which are sensitive to perceptual cues and are controlled by something like scripts and plans (Schank and Abelson 1977).[47]

The second position seems thus to be favorable. It still has some explanatory force.[48] In the long quote above, Lewis explicitly offers an explanation of rational agency in terms of habitual processes.

The question is whether, of the two explanations, the second is tenable. I think it is still too demanding. Lewis wants to understand the relevant habits as follows: *the behavior they elicit is the same as the behavior that would have been elicited if the agent had deliberated.* However, many of us continue to have "bad" habits while being aware of it.

So, the relevant habits required for Lewis' theory must be special. I think the guiding picture is that of rational and controllable behavior that is then learned and routinized. At first we go through the instructions and double check every step. Later, we "just do it." So understood, Lewis' proposal has some appeal.

But it still doesn't deal well with cases in which the controllability condition is not satisfied. Burge's sentimental hat-tippers (§4.3.2) are a case in point: the hat-tippers are somewhat stubborn or sentimental since they would rather stick to tradition than change their behavior. Insofar as we're inclined to say that there can be conventions in such scenarios – and I think we should say so –, the habits do not satisfy Lewis' controllability condition.

So, if we want to defend Lewis' proposal, we should at least endorse the position according to which agents act rationally on the basis of deliberation *or* act so as a result of controllable routinization. But both the rationality of behavior and controllability of routines are substantial assumptions which

[47] Research in experimental economics supports the hypothesis that there are other behavior-controlling mechanisms (Camerer 2003:24). Meanwhile, also philosophers have thought that conventional behavior is habitual, *e.g.* Kemmerling (1976:123) and Savigny (1985).

[48] Contra Kemmerling (1976:123).

are not trivially met and are problematic in certain cases.

4.3.5 Interim summary

In the previous sections, I've suggested some clarifications and changes: (i) The regularities in action and belief can be extended to ones in action and attitude (§4.1). This will be useful for Lewis' Signaling Games theory. (ii) Lewis' reason to explicate common interest in terms of proper coordination equilibria, rather than the more standard strict Nash equilibria, rests on shaky intuitions (§4.2.3). I don't find it necessary to follow Lewis here and suggest to explicate common interest in terms of strict Nash equilibria. (iii) Conventional "regularities" should be understood as *patterns of activity.* (iv) The common-knowledge assumption can and should be dropped. (v) Alternatives are historically accidental and do not require that the agents could change their behavior now.

Let us turn now to some objections from the literature before we check Lewis' account against the adequacy conditions from §1.4.

4.3.6 General points about the analysis

The complexity of Lewis' analysis has made many suspicious.[49] But, generally speaking, even if an analysis is bewilderingly complex, its complexity is by itself not a *good* argument against it, as long there is no rivaling account which is at least equally good and less complex. Pragmatically speaking, bewildering complexity might still be a reason not to use it if the "costs" outweigh the "benefits".

Lewis' account is idealized: There we have perfect conformity, ideally rational reasoners working out the consequences of their behavior while replicating each others' reasoning. Isn't that an argument against it?[50] This criticism misses the target since Lewis weakened the demanding assumptions quite a bit and if we follow my suggestions, then we can weaken them even more.

Another common objection is that Lewis conventions are not normative

[49] See for example Millikan (1998), Nolan (2005:158 ff.), and Schwarz (2009:189 ff.).

[50] For example Savigny (1988:§§11–15) criticizes Lewis' analysis for taking for granted that human agents are rather individuals acting rationally than social beings and thereby undervaluing the role of sanctions.

in an interesting sense.[51] Since convention-conforming behavior is rational and is explained (or at least justified) in terms of the instrumental principle, only prudential *oughts* are entailed. That is, Lewis conventions involve *oughts* with a recommending but not – as some require[52] – a demanding character. Since I distinguish between conventions and social norms (§1.2), I think that Lewis' analysis is as it should be in this respect. This is not to deny that there is also an interesting notion of a *normative convention*; in fact I introduce it in §8.2.3.

4.3.7 Conventions without common interest

According to Lewis, conventions serve a *common interest* (Lewis 1975:5) (in the sense that everyone wants to conform and wants the others to conform; see §4.3.7). But it seems that there are conventions which have other social functions. Let us consider the following cases which seem problematic for Lewis' analysis: (i) to hold the fork in the right hand, (ii) to open doors for women, and (iii) to dress as the others do.[53] These cases seem to be ordinary conventions. But in all these cases, the respective convention does not seem to exist because it serves a common interest but because people do not want to be non-conformists.

There are two striking features about these cases. First, there are normative expectations present in all the cases. For this reason, I think we should count the cases as social norms. This illustrates that the distinction between conventions and social norms is helpful: we don't have to squeeze everything into the same bag.

Second, if one follows my suggestion to understand what it is to be in the common interest in terms of strict Nash equilibria (or in terms of Nash equilibria that satisfy some stability condition), then the above cases can be counted as conventions. For example, in the "dress as the others do" case everyone would be worse off if one were to deviate. That is to say that dressing as the others do is a strict Nash equilibrium. Hence, these examples can be counted as normative conventions since they are a combination of a social norm and a convention (see 8.2.3).

[51] Also Lewis (2002:97) observes this.

[52] *E.g.* Gilbert (1989:§4.6 ff.), Kemmerling (1976:§4.3), and other proposals following Hart's analysis of a rule (Hart 1997): Savigny (1988) and Glock (2010).

[53] Cases (i) and (ii) are from Schiffer (1972:152); case (iii) is adapted from Gilbert (1989:§IV 4.4). The point is also discussed in (Schwarz 2009:192).

4.3.8 Adequacy of Lewis' account

Lewis' account comes close to being adequate according to the adequacy conditions proposed in §1.4. To highlight some important points: With regard to the pre-theoretic characterizations, Lewis' account does not include dispositional conventions (DesC1). On the basis of the *expected-utility* approach, it allows for conventions in a community with widespread nonconformity. This is required for being relatively robust (DesC1). The main issue is, in my opinion, that Lewis takes for granted that humans act rationally on the basis of rational deliberation (or, if habits took over, then they still must be under rational control). Insofar as humans function differently, the account does not satisfy the condition that conventional behavior must be feasible for humans (DesC5). I think that this is indeed so. So one should aim at an account that allows that agents conform to conventions on the basis of social habits and other non-deliberating mechanisms.[54]

4.4 Summary

In this chapter, I've discussed Lewis' account of conventions and rehabilitated his game theoretical definition. Lewis oscillates between an analysis of rationalistic conventions and rationally justifiable conventions, while dispositional conventions are not considered.

I've argued that Lewis' analysis should receive six changes: (i) The regularities in action and belief can be extended to ones in action and attitude (§4.1). (ii) "Regularities" should be understood as *patterns of activity* (§4.3.3). (iii) The almost perfect conformity condition should be weakened to "often enough" conformity (§4.3.3). (iv) The candidates for conventions and alternatives should be restricted to those that are feasible (§4.3.2). (v) We should explicate common interest in terms of strict Nash equilibria (§4.2.3). (vi) The common-knowledge assumption should be given up (§4.3.1).

[54] For similar verdicts, see *e.g.* (Burge 1975), (Kemmerling 1976:§1), (Grandy 1977), (Savigny 1988:§§11–15), (Camerer 2003:24).

Chapter 5

Signaling Games

> The International Code of Signals lists a correspondence of flag hoists and certain predicaments. That is, it gives a contingency plan for ships (strictly, for ships' officers) of the form: if in such-and-such predicament, hoist such-and-such flags. There is a complementary contingency plan for ships that observe flags on nearby ships: if a ship hoists such-and-such flags, act as would be appropriate on the assumption that it is in such-and-such predicament. Ships do regularly act according to these two complementary contingency plans.
>
> *Convention*
> David Lewis

The chapter epigraph illustrates the Signaling Games approach: There is a conventional signaling regularity consisting in senders giving signals and receivers reacting to them. If such a convention prevails, the signals are said to have a meaning.

In philosophy Signaling Games go back to David Lewis. Lewis rejected his theory since he thought it was too limited to explain certain features of human languages (§5.4). Nevertheless, we shouldn't dismiss such accounts too quickly. One reason is that they are simply so appealing: they are simple and instructive. Many topics of the conventionalist project can already be discussed on their basis. Moreover, researchers from different fields continue to contribute to the Signaling Games tradition.[1]

[1] For a recent overview of the theory of signaling games, see (Sobel 2009). For recent overviews in philosophy and in linguistics, see (Rescorla 2007) and (Benz et al. 2005a), respectively. In biology, the model of Grafen (1990) has been influential. In economics, the article by Spence (1973) is the classic. Kemmerling (1976:64–74) provides an early and comprehensive discussion of Lewis' theory.

I'll focus on Lewis' theory since it's the one that is philosophically relevant. Lewis' main aim in developing the account was to refute two spectacular claims of Quine. First, following Russell, Quine (1976a) challenged conventionalist analyses of meaning by pointing out that conventions are something themselves requiring a language. That is, Quine invoked the regress argument I mentioned in §1.1.4. Second, Quine (1980) argued in his classic *Two dogmas of empiricism* against the "dogmatic" analytic/synthetic distinction. Lewis accepted Quine's arguments as a challenge and proposed a solution to them in *Convention* (Lewis 2002). The solution consists in two parts. First, Lewis gives a careful analysis of a notion of a convention which does not depend on language – the one we know from the last chapter. Second, he analyzes a notion of meaning which can be used to defend the analytic/synthetic distinction in terms of Lewis conventions. The basic model he uses for the second part is a so-called "two-sided signaling problem" or "signaling game," as it is called today.

If, in a signaling game, agents give signals and react to them in a certain conventional way, we say that there is a *signaling system*. This is deliberately left vague at the moment but with it in mind, we can characterize a Signaling Games account by the following three claims:

SG1. A theoretically interesting part of natural languages can be explained by signaling systems.

SG2. Important parts of a signaling system are: (i) signals, (ii) a population, (iii) a sender and receiver role, (iv) states of affairs observable by senders, (v) reactions of receivers, (vi) a pattern of activity prevailing in the population to use *signals* according to certain *contingency plans*. The contingency plans for the members of the population consist of two parts, one for the sender-role and one for the receiver-role.

a. A sender's plan determines which signal to produce depending on what the sender has observed, her desires, and her beliefs.

b. A receiver's plan determines how to react to an observed signal depending on her desires and beliefs.

SG3. Signals in a signaling system have a meaning in virtue of the fact that there is a pattern of activity that is a convention in the population. Which meanings the signals have depends on the pattern which is the convention.

The meaning of these claims will become clear as we go on. But, let me quickly elaborate on what I mean by SG1. What I have in mind here is what Resnik (2002:4) calls an "explanatory idealization." An explanatory idealization provides a model for several aspects of a phenomenon which are deemed as important and analyzes the relations of the modeled parts. The model is explanatory insofar as it has explanatory force for the modeled slice of reality.

The plan for this chapter is as follows. I begin by describing three motivating examples of signaling games in §5.1. Thereby I informally introduce the core ideas of Signaling Games accounts. In §5.2 I present Lewis' Signaling Games theory. I discuss it in §5.3; in particular, I offer a rejoinder to the regress argument in this section and consider the question whether Lewis is indeed Gricean. I'm not discussing Lewis' solution to rehabilitating the analytic/synthetic distinction. It is a contested topic and for my purposes, it is not necessary to discuss it.[2] I evaluate Lewis' theory in §5.4. The chapter ends with a summary in §5.5.

5.1 Motivating examples

Signaling is ubiquitous, ranging from speaking, gesturing, and flag hoisting among humans to various behaviors among animals. I'll focus on homely cases of signaling which show some variation. The variation I'm interested in relates to the distinction between rationalistic, rationally justifiable, and dispositional conventions (§1.2). I provide for each case an example of signaling.

5.1.1 Newman's lanterns and Revere's warnings

The following true story is Lewis' famous example of a signaling game.[3] It's an example of signaling on the basis of rational deliberation.

Paul Revere, an American war activist during the Revolutionary War, instructed Robert Newman, the sexton of the Old North Church of Boston, to inform the colonists in Charlestown about the movements of the British troops (who were called "the redcoats" in those days) which Newman could observe. Revere and Newman were prepared for three situations: The redcoats staying home, their setting out by land, and their setting out by sea.

[2] But see (Lewis 2002:173–177, 195–202, 203–208) for Lewis' discussion.

[3] See (Lewis 2002:122 ff.). Two well-sourced Wikipedia articles about Revere and his communication with Newman are (Wikipedia 10.03.2009) and (Wikipedia 13.03.2009).

If the redcoats were staying home, Newman would go home. If they set out by land, Newman would warn the colonists that the redcoats are coming by land. If they set out by sea, Newman would warn the colonists that the redcoats are coming by sea. Revere's and Newman's joint problem was to communicate without being noticed by the redcoats. This turned out to be difficult since riding from Boston to Charlestown was too risky. Mind you, this was 1775: no Morse code, no telephones, no Internet. Revere solved their problem by instructing Newman to communicate by using lanterns as signals. But how did Newman and Revere manage to communicate?

Newman acted according to a contingency plan. An example of a contingency plan is the following one (called "S1"):

- Hang no lantern, if the redcoats are observed staying home.
- Hang one lantern, if they set out by land.
- Hang two lanterns, if they set out by sea.

Newman could have acted otherwise. There are many other contingency plans which only involve these three signals (no lantern/one lantern/two lanterns) and the three conditions (redcoats stay home/set out by land/set out by sea). To mention but another one of them ("S2"):

- Hang no lantern, if the redcoats are observed setting out by sea.
- Hang one lantern, if they stay home.
- Hang two lanterns, if they set out by land.

Also Revere acted according to a contingency plan ("R1"):

- Go home, if no lantern is observed.
- Warn the countryside that the redcoats are coming by land, if one lantern is observed.
- Warn the countryside that the redcoats are coming by sea, if two lanterns are observed.

Likewise, also Revere could have acted otherwise since there are many other contingency plans for him. Revere and Newman had a common interest in successful communication. But not every combination of individual contingency plans is suitable to achieve this. *E.g.*, the combination S2-R1 would have disastrous consequences when the redcoats set out by sea: Newman would hang no lantern and upon this observation Revere would go home, leaving the countryside unwarned. So, to communicate successfully, Newman and Revere have to coordinate. Historically, communication between Newman and Revere was successful. They acted according to a *signaling convention*, *i.e.*, to act according to such contingency plans. Newman acted according to S1 and Revere according to R1. They could also have coordinated on another pair of "suitable" contingency plans (*e.g.* by swapping the signals in a suitable way). But for each of them it was optimal to act the way he acted on condition that the other stuck to his contingency plan.

Being a good story, what's its moral? After all, we wanted to have a conventionalist theory of meaning but in no word was "meaning" mentioned. However, this is as it should be. We want to have a description of those non-semantic facts on which the semantic facts supervene; or at least, the description should not use the notion of linguistic meaning, while it may use more basic semantic notions. Lewis' insight was that in the case of Newman and Revere we are inclined to attribute meanings to the signals they use. In other words, Lewis proposed that if Newman and Revere do signal and act regularly in a conventional way where both act according to suitable contingency plans, then the signals have meanings:

> I have now described the character of a case of signaling without mentioning the meaning of the signals: that two lanterns meant that the redcoats were coming by sea, or whatever. But nothing important seems to have been left unsaid, so what has been said must somehow imply that the signals have their meanings. (Lewis 2002:124–125)

According to Lewis' Signaling Games proposal, the signaling convention determines the meaning of the signals – just as the conventionality thesis has it. But, in this case, Newman instructed Revere and thereby they were using a language they already had to establish their signaling convention. In general, we don't want to assume that signaling conventions depend on a prior shared language. Fortunately, we don't have to.

5.1.2 Suske's weather forecasts and Wiske's dress

A simpler example without presupposing language and rational deliberation is this. Suske and Wiske have, for whatever reason, a common interest in the following: Whenever it will rain on the coming day, both want Wiske to take a raincoat with her. And whenever it will be sunny on the coming day, both want Wiske to wear her beautiful sunglasses. Depending on the weather report Suske hears, he gives her a certain signal. When it's going to rain, he shows her a green card. When it's going to be sunny, he shows her a red card. Wiske reacts to seeing a green card by taking a raincoat with her on the coming day. She reacts to seeing a red card by wearing sunglasses. This is a regularity between Suske and Wiske, developed habitually over time. In the beginning, coordination was not so successful but over time, they conformed more and more. Nowadays, it's a rationally justifiable convention between them.

Again the notions of meaning and other semantic notions in the neighborhood were nowhere used in the description of the example. Nevertheless, I take us to be inclined to say that showing a green (red) card *means* something among them. There are two obvious meaning attributions. First, showing a green (red) card means *that it's going to rain (be sunny) on the coming day* among them. Second, showing a green (red) card means *to take a raincoat with you (to wear sunglasses) on the coming day* among them.

5.1.3 The alarm calls of vervet monkeys

The last two examples made an implicit assumption, namely that the agents act rationally in the sense that they choose their actions or contingency plans by deliberating. In the first example, Newman and Revere are assumed to deliberate. In the second example, Suske and Wiske didn't deliberate but they could have. But is deliberating or the possibility to do so a strict requirement? I think not. The signaling behavior of vervet monkeys is a case in point where we have signaling on the basis of behavioral dispositions but presumably without rational deliberation.

Vervet monkeys use different alarms calls to warn their peers about predators. They are reported to have three alarm calls, "pyow" for leopards, "hack" for eagles, and a third one for snakes (whose sound never seems to be described in the literature). If a vervet observes one of these predators, then it gives a suitable alarm call. That is, if a vervet observes a leopard, then it produces "pyow-pyow-..." sounds. If it observes an eagle, then it produces "hack-hack-..." sounds, and so on. Upon receiving one of these calls, the peers show evasive behavior. The evasive behavior is specific to the alarm call. For example, upon hearing "hack," vervets typically freeze because eagles pick up their position by observing their movements. But the evasive behavior associated with an alarm call is not always exactly the same. The vervets do what is best for escaping in the very situation they find themselves in. For this reason, it seems more apt to say that an alarm call *indicates that the respective predators are invading the territory* (in the sense of *being a natural sign of*) than to say that an alarm call *instructs the* *peers to escape in a certain way.* Moreover, the vervets' signaling behavior seems to be the result of an evolutionary adaption and there seems to be no evidence that vervets consciously and deliberately choose to signal and react in the ways they do.[4]

5.2 Lewis' Signaling Games theory

The three examples considered in the last section have a common structure: First, the examples consist of a certain type of strategic situation in which signals play an important role to coordinate the agents' behavior. Second, there is a certain pattern in their behavior. Third, the signals used in the respective situations can be attributed a content.

Lewis' proposal is that such contents of signals are their meanings if the pattern of activity is a *signaling convention.*[5] According to the proposal, the first part of the common structure is analyzed in terms of so-called *signaling games*, a class of games in which the strategic role of informational asymmetry is studied. The second part is analyzed in terms of Lewis conventions and the third relates to a meaning-determination claim.

Lewis calls the kind of signaling game he uses "two-sided signaling problems" (see Lewis 2002:130 ff.). I'll use Lewis' term to avoid misunderstandings.[6]

By using *two-sided signaling problems*, Lewis' theory emphasizes the

[4] A classic reference about the signaling of vervet monkeys is (Cheney and Seyfarth 1992) but see also the blog article by Lieberman (26.05.2006) referencing and evaluating more studies.

[5] Surprisingly, Lewis (2002:147) remarks that conventions are not necessary to "define meaning for signals." I'm not sure what Lewis' point is. Maybe the point is that to define what a signal means in a possible signaling system, the signaling system does not have to be actually used. Or maybe he wanted to allow for signaling systems that are used among vervet monkeys who are, arguably, cognitively too limited to be party to a Lewis convention. If the latter, I think he should have chosen a notion of a dispositional convention. Subsequently, I ignore Lewis' remark.

[6] Two remarks are in order. (i) In terms of the theory of signaling games, a two-sided signaling problem could be called more precisely a *deterministic one-sender/one-receiver signaling game with costless signals without meaning* using proper coordination equilibria as the solution concept. (ii) Lewis (2002:128–130) also considered one-sided signaling problems in which either only senders or only receivers in a population coordinate, *e.g.* horse-riders who guide their horses by yelling "gee!" and "haw!" – where the horses don't coordinate with the riders but the different riders do, to train the horses to react in a certain way.

strategic aspects of signaling and thereby of communication in general. For example, it describes the situation of Suske and Wiske in a slightly different way:

(1) Suske has access to weather reports and, based on this information, he sends a signal to Wiske. Wiske, in turn, has no access to the weather reports but does to Suske's signals. Based on what Wiske saw, she reacts by taking the appropriate dress with her. Each of them has preferences over the way Wiske's reactions are related to the state of the world Suske observed. To serve their individual interests, each of them has to choose among the contingency plans (or strategies) available to them.

5.2.1 Two-sided signaling problems

A *two-sided signaling problem* is a type of situation with two agents, S ("sender") and R ("receiver"), a finite set T ("types") of states the world can be in, a finite set M of messages (or signals) S can send to R, a finite set A of actions R can perform upon having received a message, and for each agent payoff functions U_S and U_R from $T \times A$ into $\mathbb{R}$. The strategy spaces are defined as follows. A strategy σ of S is a function from T into M, interpreted as acts of signaling messages m depending on the states t of the world. A strategy ρ of R is a function from M into A, interpreted as reactions a depending on the observed message m. A strategy profile $\langle \sigma, \rho \rangle$ of a two-sided signaling problem is any tuple consisting of a sender strategy σ and a receiver strategy ρ. Formally, two-sided signaling problems can be modeled by tuples of the form $\langle \{S, R\}, T, M, A, U_S, U_R \rangle$.

The following assumptions are made about the two-sided signaling problem: S can reliably observe the state $t \in T$ the world is in, *i.e.*, S is assumed to know t (or at least to act as if she knew it). S has a certain preference U_S over R's actions $a \in A$ depending on the state t the world is in. S can send a message $m \in M$ to R (by signaling m to R). Upon observing the state the world is in, S sends some message $m \in M$ to R. R also has certain preferences U_R over her reactions which also depend on the state of the world. R cannot observe the state of the world and is uncertain about the actual state. R can reliably observe S's signals. Upon observing S's message m, R reacts to it by performing an action $a \in A$. Furthermore, the payoff functions of the sender and receiver are assumed to have the following properties. (i) If in a state t the receiver reacts in a way preferred by the sender, the payoffs $U_x(t, a)$ are 1, for $x = S, R$; otherwise

0. (ii) the payoff functions U_S and U_R are such that they determine a unique dependency F of the receiver's behavior on the state t the world is in that both prefer to other possible dependencies.[7] F can be described by a function from T into A. For simplicity, we assume that the states the world can be in are equiprobable (*i.e.* $p(t) = 1/|T|$) and that the agents only care about whether R's reaction a in some state t accords with action $F(t)$ or not. The payoff functions U_S and U_R for the strategy profiles of the sender and the receiver are then defined as expected utilities, *i.e.*: $U_x(\sigma, \rho) = \sum_{t \in T} p(t) \times U_x(t, \rho(\sigma(t)))$. Since, the states are equiprobable, we can simplify the definition to $U_x(\sigma, \rho) = \sum_{t \in T} U_x(t, \rho(\sigma(t)))$. Finally, T, M, and A are assumed to have the same number of elements.[8]

A consequence of this definition is that in a two-sided signaling problem, both agents face an interdependent decision problem. S can choose which message to signal based on her observation of the actual state of the world. R can choose how to react to the message she observes. The strategic problem is that both want that R's reactions depend according to F on the state the world is in. However, while S knows the state the world is in, S cannot make R to act accordingly but can only signal some message. And while R can act in certain ways, R does not know the world is in.

5.2.2 The two-sided signaling problem of Suske and Wiske

We can apply the definition of a two-sided signaling problem to the Suske-and-Wiske example. It illustrates the simplest non-trivial case, namely a two-sided signaling problem G_1 with two agents, two states $\{\text{rainy}, \text{sunny}\}$, two messages $\{\text{green card}, \text{red card}\}$, two actions $\{\text{wear sunglasses}, \text{take raincoat}\}$, and a coinciding interest in coordination, as the payoff table over the product of the states and actions shows:

[7] If S and R's preferences are so unaligned that there is no such F, then, by stipulation of what a two-sided signaling problem is, they can't be part of a two-sided signaling problem.

[8] These definitions are slightly more restrictive than the one from Lewis (2002:130 ff.): (i) Lewis allows for an audience consisting of more than one member, *i.e.* for one-sender/multi-receiver-scenarios. (ii) Lewis only assumes that $|T| = |A| \leq |M|$. (iii) Lewis does not require that the states $t, t' \in T$ are equiprobable. (iv) Lewis does not use the simple utilities over states and actions but only requires a common interest in a certain state-reaction-dependency which I designate by "F". For our discussion, we don't need the more general definitions since the points I make also go through in the more general case.

	To wear sunglasses	To take raincoat
It's rainy	$0, 0$	$1, 1$
It's sunny	$1, 1$	$0, 0$

The common interest in coordination determines a dependency F of Wiske's reactions on the states as follows:

- It's rainy ⇒ Wiske takes her raincoat.
- It's sunny ⇒ Wiske wears sunglasses.

That is, if it is sunny, both desire that Wiske reacts by wearing sunglasses. If it is rainy, both desire that Wiske reacts by taking her raincoat. In the two-sided signaling problem they are in, Suske consequently should choose the messages *green card* and *red card* in the states she observes so as to achieve coordination. This in turn requires that Wiske reacts to the messages appropriately. The strategic situation is as follows. Suske's strategies are:

- GG: to always show the green card,
- RR: to always show the red card,
- GR: to show the green card when it's rainy and to show the red card when it's sunny, and
- RG: to show the red card when it's rainy and to show the green card when it's sunny.

Wiske's strategies are:

- SS: to always wear sunglasses,
- RR: to always take her raincoat,
- SR: to wear sunglasses when observing the green card and to take her raincoat when observing the red card, and
- RS: to wear sunglasses when observing the red card and to take her raincoat when observing the green card.

Applying the expected utility definition, the following strategic game results:

	SS	SR	RS	RR
RR	$1, 1$	$1, 1$	$1, 1$	$1, 1$
RG	$1, 1$	$2, 2$	$0, 0$	$1, 1$
GR	$1, 1$	$0, 0$	$2, 2$	$1, 1$
GG	$1, 1$	$1, 1$	$1, 1$	$1, 1$

5.2.3 Admissible strategies and signaling systems

The simplest two-sided signaling problem illustrated by Suske and Wiske makes us aware of the fact that already simple cases lead to a complex strategic situation. Since both Suske and Wiske have each 4 strategies, there are already 16 strategy profiles in this simple case.[9] The strategy profiles fall into three categories. First, there are the ones like $\langle \text{RG}, \text{RS} \rangle$ and $\langle \text{GR}, \text{SR} \rangle$ in which the agents manage to always miscoordinate. In such a situation, Suske does not get across what he wants to signal since Wiske does not react in a way realizing the commonly desired dependency F. In these cases, there is some sort of communication but it's miscommunication. Second, there are strategy profiles like $\langle \text{RR}, \text{SS} \rangle$ where the receiver at least sometimes reacts in a way desired by the sender. Here, there is no communication because the receiver's action is independent of the signal sent. Third, there are strategy profiles like $\langle \text{RG}, \text{SR} \rangle$ and $\langle \text{GR}, \text{RS} \rangle$ which maximize the expected utilities because in all states coordination is achieved. That is, they realize communication. The agents' coordinated actions accord with the commonly desired dependency F.

Lewis singles out the strategy profiles of the third category as follows. A sender's strategy σ is called "admissible" iff it is a one-to-one function from T into M. A receiver's strategy is called "admissible" iff it is a one-to-one function from M into T. By restricting the strategy profiles to those consisting of admissible strategies, we arrive at the following restricted strategic game:

	SR	RS
RG	2, 2	0, 0
GR	0, 0	2, 2

The strategy profiles of the restricted game now either belong to the *miscoordination*- or to the *coordination*-category. Moreover, the strategy profiles $\langle \sigma, \rho \rangle$ bringing about coordination combine so as to realize F: in all states t the world can be in, the receiver's reaction to the sender's signal is the same as the one determined by F, *i.e.* $\forall t \in T : \rho(\sigma(t)) = F(t)$. Lewis calls such strategy profiles "signaling systems." It is easy to see that signaling systems are proper coordination equilibria (see §4.2.3) in the strategic

[9] In general, in a two-sided signaling problem with $|M|$ messages, each player has $|M|^2$ strategies yielding a space of strategy profiles of size $|M|^4$.

game restricted to admissible strategies. *Proof*: Assume that $\langle\sigma,\rho\rangle$ is a signaling system. Then, by definition of the utility functions, $\langle\sigma,\rho\rangle$ maximizes the payoff of U_x for $x = S, R$ since in every state t, $\rho(\sigma(t)) = F(t)$. Alternatives to $\langle\sigma,\rho\rangle$ are $\langle\sigma',\rho\rangle$ and $\langle\sigma,\rho'\rangle$ for a sender's admissible strategy profile $\sigma' \neq \sigma$ and a receiver's strategy profile $\rho' \neq \rho$. But alternatives do not maximize the payoff since for at least one element in the domain of σ' or ρ', respectively, the function yields a different value such that for t, $\rho(\sigma'(t)) \neq F(t)$ (or $\rho'(\sigma(t)) \neq F(t)$, respectively). Thus, since $\langle\sigma,\rho\rangle$ is a row- and column-maximum, $\langle\sigma,\rho\rangle$ is a proper coordination equilibrium.

This should remind us of Lewis' analysis of conventions. For coordination problems were analyzed in terms of proper coordination equilibria. It follows from the results Lewis has proved (Lewis 2002:132 ff.) that every non-trivial ($|M| > 1$) two-sided signaling problem has at least two proper coordination equilibria. Thus, non-trivial two-sided signaling problems are a kind of coordination problem as introduced in §4.2.3.

5.2.4 From signaling conventions to meaning

That two-sided signaling problems relate to conventions is not accidental. We can make the relation more explicit. A signaling convention is a (Lewis) convention applying to two-sided signaling problems, understood now more abstractly such that S and R denote *roles* instead of particular agents.[10] That is, signaling conventions are conventions applying to a special type of situation. A little bit more precisely, a *signaling convention with regard to a two-sided signaling problem G among a population P* is a convention among P whose members are involved in pairs as senders and receivers in the two-sided signaling problem G in which they behave according to a strategy profile $\langle\sigma,\rho\rangle$ which is a signaling system. We then say that $\langle\sigma,\rho\rangle$ is a *conventional signaling system among P with regard to G*. Signaling conventions have a goal or social function (§4.3.2): They guarantee that senders and receivers behave so as to realize the commonly desired dependency F of G. In other words, signaling conventions are a means to realize the dependency F of an abstract two-sided signaling problem in a population.

Suske and Wiske's pattern of activity is a first example of a signaling convention. In the notation introduced above, they behave according to the

[10] This kind of move to abstract games from the particular agents has been discussed in §4.2.4.

strategy profile $\langle \mathrm{GR}, \mathrm{RS} \rangle$. The profile is a signaling system and a convention among them. Let us reflect on what they do. Suske shows the green card when it's rainy and the red one when it's sunny. Wiske reacts by taking her raincoat or wearing sunglasses, respectively. So, we can say that showing the green card indicates both that it's rainy and that Wiske is going to take her raincoat. It is now not without appeal to say that showing the green card *means* that it's rainy. And likewise, that showing the green card *means* that Wiske is to take her raincoat. (Hence the message "showing the green card" is ambiguous; I'll return to this below in this section and in the subsequent section.)

This reflection on what Suske and Wiske do by being party to their signaling convention suggests that conventional signaling patterns determine the meaning of the signals. There is even a simple procedure to determine the meanings of the signals in a signaling convention.

Observe that a signal has two meanings. One of them is linguistically expressed by a "that"-clause, the other by a "to"-clause. In the running example, the first meaning of showing the green card is *that it's rainy.* The second meaning is *to take her raincoat.* So, the first meaning designates what is the case while the second meaning designates what is to be done. Consequently, let us call the first meaning the signal's *indicative meaning* and the second meaning its *imperative meaning.* If a signaling convention to act according to some strategy profile $\langle \sigma, \rho \rangle$ prevails among a population in a two-sided signaling problem $\langle \{S, R\}, T, M, A, U_S, U_R \rangle$, then, for $t \in T$ and $m \in M$, the indicative meaning *that t is the case* of a message m is determined by the pre-image of the sender-part of the pattern as follows. The indicative meaning of m in these circumstances is *that the single state t in $\sigma^{-1}(m)$ is the case.* Since σ is one-to-one, it is guaranteed that there is exactly one element in its pre-image. The imperative meaning *to do a* of a message m is determined by the image of the receiver-part of the pattern as follows, for some $a \in A$ and $m \in M$. The imperative meaning of m in these circumstances is *to do action $a = \rho(m)$.* Since ρ is also one-to-one, it is guaranteed that there is exactly one element in its image.

5.2.5 From signals to sentences

The account developed so far can be generalized. Until now, we've only considered the simple two-sided signaling problem G_1 between Suske and Wiske with two messages which could be used to perform communication acts somewhat similar to speech acts of informing and commanding. But one might well ask how signaling messages relates to linguistic conversation.

The current account lacks several linguistic core notions: language, linguistic expression, grammar, and a proper semantics. In particular, the messages have no syntactic structure. For this reason, it's a stretch to call them "sentences." Moreover, there is neither a notion of *truth in a language*, nor are important semantic phenomena like reference, indexicality, and vagueness taken into account. Also, the relation of the messages to their use is simplistic. At least one central speech act, the one of asking, is omitted, and clearly, there are many more.

As Lewis (2002:141–152) has shown, some of these concerns can be addressed easily. The step to truth is simple. If a signaling convention prevails, then we can define a message's truth condition relative to a situation as follows. Let there be a signaling convention to do $\langle \sigma, \rho \rangle$ in a two sided signaling problem $\langle \{S, R\}, T, M, A, U_S, U_R \rangle$ among a population P. Then for a message $m \in M$, according to its indicative meaning *that t is the case*, m is true in an utterance situation s iff in s, t is the case. *E.g.* the green card is true in an utterance situation s iff in s, it is rainy.

From the example we know that messages can be assigned different kinds of meanings, at least indicative meanings and imperative meanings. Do we have to decide between the two for assignments of truth conditions and if so, how? Does it make sense to say that a message is "true" in its imperative meaning? With regard to the latter question, we can extend our talk to *satisfaction conditions.* In this sense, signals in either meaning can be satisfied in an utterance situation. The former question is trickier. Lewis wanted to decide between the two as follows. If the receivers in a two-sided signaling problem deliberate upon receiving the signal before they act, then this is evidence for assigning the signal an indicative meaning. But if they just react to the signal, then it has an imperative meaning (*cf.* Lewis 2002:144, 146).[11]

Given a signaling convention and for each message a mood assignment, we can derive a function $\mathcal{L}$ which assigns to every message a complex mean-

[11] But see (Millikan 1995b, 2006) and in particular (Harms 2004) and (Huttegger 2007) who argue that we needn't decide between the two meanings. For a signal can have both meanings: It can indicate that something is the case and demand or motivate to do something. Their proposals are discussed in chapter 7 (without focusing on this special topic).

ing $\langle \mu, \tau \rangle$ where μ is the signal's *mood* and τ is its *satisfaction condition.*[12]

The two components of a complex meaning can be interpreted as follows. A mood can be identified with the signaling act's *illocutionary force.* A satisfaction condition for indicatives is a truth condition; for other moods it's (typically) something else, *e.g.* for imperatives it's the condition that almost every receiver obeys the imperative meaning. Moods are abstract entities. Lewis uses code numbers: 0 for indicative and 1 for imperative.

In the simple two-sided signaling problem between Suske and Wiske, the red card's indicative meaning is now coded as the pair $\langle 0, \textit{that it is sunny} \rangle$ and its imperative use as $\langle 1, \textit{to wear sunglasses} \rangle$. If we have evidence that the red card is used imperatively, then we set the value of $\mathcal{L}$(red card) to $\langle 1, \textit{to wear sunglasses} \rangle$; otherwise we set it to $\langle 0, \textit{that it is sunny} \rangle$.

If, in addition, the messages are now *verbal expressions* and the sender's signaling acts are acts of uttering such expressions, then the two-sided signaling problem is more akin to a model of linguistic communication. Verbal expressions are finite sequences of types of sounds or marks which satisfy the following conditions: (i) they can be tokened by the senders (by uttering or inscribing them), (ii) their tokens can be easily observed by the receivers, (iii) no party has strong (extraneous) preferences for or against them being tokened and the preferences are independent of the state the world is in.[13] Lewis' called such two-sided signaling problems which have verbal expressions as messages "verbal."

We can now define what is for a verbal expression to be satisfied in $\mathcal{L}$ in an utterance situation. The definition is relative to a signaling convention of a verbal two-sided signaling problem among members of a population P in which the signal meaning of any $m \in M$ is $\mathcal{L}(m) = \langle \mu_m, \tau_m \rangle$. We say that an utterance of a verbal expression m is "satisfied in $\mathcal{L}$ in the utterance situation s" iff τ_m is μ_m-satisfied in s. A satisfaction condition τ is μ-satisfied in s as follows. If $\mu = 0$ (indicative), then τ is μ-satisfied iff τ in s is the case. If $\mu = 1$ (imperative), then τ is μ-satisfied in s iff

[12] The idea to describe the signal meanings in this way is based on Stenius (1967).

[13] *Cf.* Lewis (2002:141 ff.). I think that these conditions should only be understood as *indicators* for some behavior to be verbal signaling behavior. Savigny (1988:§19) proposes some more indicators which seem plausible to me: The perception of a token depends only to a little degree on the attentiveness of the receiver, the tokens have few other functions than to be used as signals, the production of a token should be easy and not be very expensive for a sender, and they should have few exchangeable parts, *i.e.* their parts make a specific contribution in the whole signaling-pattern.

what almost every receiver in s does is τ.[14] If the set of messages are verbal expressions having a meaning assigned by a function $\mathcal{L}$, $\mathcal{L}$ can be said to be a *signaling language* and the verbal signals can be called "sentences." The signaling language might be rudimentary but this is no reason not to call it a language.

There are two noteworthy points. First, we had to add an index s for the utterance situation to the assignments of satisfaction conditions. Having such an index allows us to generalize the setting to signals with an indexical meaning in the obvious ways. *E.g.* the indicative verbal expression "I am here" could have the following satisfaction condition: the sender in s is in s at the location of s.

Second, we can use the foregoing to say what it is for a population P (or a linguistic community, if you like) to use a signaling language $\mathcal{L}$. We can do so in two steps: (i) We have to define what Lewis (2002:148 ff.) calls a "convention of truthfulness." (ii) We use this definition to define what it is to use a language.

If $G = \langle\{S, R\}, T, M, A, U_S, U_R\rangle$ is a two-sided signaling problem with a strategy profile $\langle\sigma, \rho\rangle$ of G and P is a population of agents, then we say that "there is a convention of truthfulness in $\langle\sigma, \rho\rangle$ among members of P" iff $\langle\sigma, \rho\rangle$ is a conventional signaling system among P with regard to G. This amounts to saying that senders try to utter whichever verbal expression is satisfied in that instance, and the receiver responds by doing whatever seems best on the assumption that the sender has succeeded in uttering a verbal expression which is satisfied.[15]

Then, we can say that "a population P uses a signaling language $\mathcal{L}$" iff

1. there is a two-sided signaling problem $G = \langle\{S, R\}, T, M, A, U_S, U_R\rangle$ with a strategy profile $\langle\sigma, \rho\rangle$ of G, and
2. there is a convention C of truthfulness in $\langle\sigma, \rho\rangle$ among members of P, and
3. $\mathcal{L}$ is the signaling language determined by C.

Let us quickly look back before we move on. The generalized theory could make up for several of the linguistic core notions we missed. One

[14] I deviate here in style but not in substance from Lewis since I want to prepare a generalization to add more moods. See §5.3.2.

[15] That is, receivers could be called "trusting" since it is assumed that they believe that senders are truthful and act accordingly. Their assumption is generally justified since the existence of a convention entails that the defining clauses of the conventions are satisfied, in particular, the common knowledge condition.

of the remaining limitations are the missing moods. For questions, Lewis proposed that we should treat them as special imperatives of the kind: "I want you to tell me ...," *cf.* (Lewis 1970). But can we add more moods for other illocutionary forces? I think we can, see below. Another limitation is that only *sentences* are represented in verbal two-sided signaling problems. The reason is that messages are *unstructured.* So, we also lack a grammar and something which assigns a meaning to words and other kinds of subsentential expressions.

Additionally, it seems implausible that his theory can explain the signaling behavior of vervet monkeys. The reason is that it is implausible that there can be a Lewis convention among vervet monkeys because (very likely) these animals couldn't rationally deliberate as Lewis requires.[16] (The other two cases – Newman/Revere and Suske/Wiske – can readily be explained.)

5.3 Discussion

We can now explicate the "use determines meaning" slogan for Lewis' Signaling Games theory. To do so, we have to answer three questions which relate to the form of the slogan. To make a determination claim, one has to elaborate on the kind of determination relation and on the relata of the relation. The determination relation in Lewis' theory can be understood in terms of *conceptual entailment* (§1.1.2): Propositions that there are signaling conventions in a population conceptually entail propositions that certain verbal expressions have certain meanings among the members of the population.

The "meanings" which are assigned to signals, and in particular verbal expressions, are complex. They are tuples of the form $\langle \mu, \tau \rangle$ consisting of a mood μ and a satisfaction condition τ. So, we could say that the meanings are *speech act types with content.* We learn little about the communicative role of these meanings since Lewis does not relate them to notions such as what-is-said or implicatures in his works on conventionalist use theories of meaning (Lewis 2002, 1975, 1976, 1992). But there is a connection between signaling in a signaling system and the sender speaker-meaning something (in a Gricean sense of "speaker-meaning") to which I turn below in §5.3.4.

[16] Millikan's theory can explain such cases since she has an account of dispositional conventions; see chapter 7.

Let us turn now to some interesting questions which are worth discussing: (i) What are the consequences of Lewis' two definitions of a convention? (ii) Can Lewis' theory be extended to other moods? (iii) How can the regress argument be escaped? (iv) Is Lewis' theory Gricean? I'll turn to these questions in this order.

5.3.1 Reconsidering Lewis' two definitions of a convention

Lewis used for his Signaling Games theory his first definition of a convention (see §4.1). This is the one whose regularities are only *in action* and not also *in action and attitude*. It's also the one which requires that almost all members of the population have almost coinciding preferences. The two limitations of the first definition of a convention effect corresponding limitations in the Signaling Games theory. I'll first state the problems for each of them and then consider a solution to it. It is based on Lewis' "official" definition of a convention from *Languages and language* (Lewis 1975), a new description of the communicative patterns in terms of speaker-meaning and understanding, and a distinction between what I call "near" and "far" preferences.

The first limitation requires that the communicative patterns must be regularities in *action*. Lewis chose a regularity in which senders signal a message upon observing a state and in which receivers react to the signal by performing some action. But defining the regularities in this way is implausible. For then there must be for each signal exactly one typical receiver reaction. But consider the following example message:

(2) "It's raining"
  a. Reaction 1: I don't leave the house.
  b. Reaction 2: I take a raincoat with me.
  c. Reaction 3: I think about the last time when I got soaking wet.

None of these three reactions seem to be atypical and it seems artificial or even outright wrong to choose one as *the* conventional reaction (moreover, the third reaction is not possible – for thinking is not an action).[17] But

[17] I use "thinking" here in the sense of *entertaining a proposition*. While entertaining a proposition is something one might cause oneself to happen, it's not an action since such events do not have intentional explanations and cannot be evaluated as rational or not; see Hunter (2003).

according to Lewis' proposal, we must commit ourselves to one reaction only.

Likewise for senders, there must be for each state exactly one message which the senders signal:

(3) The world is in the state *that it's raining*

a. Signal 1: I utter: "It's raining."

b. Signal 2: I utter: "Damn it!"

c. Signal 3: I remain silent (the empty message) while thinking about the sunny beaches of Valencia.

Arguably, it also seems wrong to claim that exactly one of the signals is *the* conventional one for the given state. The moral seems to me that in actions, we shouldn't expect that there is such a conventional pattern. The two examples are evidence for the claim that in general there are no one-to-one relations from states to messages and from messages to reactions in bigger populations with a moderately complex signaling behavior.

The second limitation is the requirement of coinciding preferences over action combinations including receiver reactions. This is equally implausible. We humans don't have coinciding preferences.[18] We are different and sometimes want to be different. Sometimes we want to manipulate others while wanting or even counting on that they understand what we say. But the possibility of conventional signaling with such manipulative desires is ruled out if it amounts to anything more than the exceptional deviations which are covered by the "almost" qualifications of Lewis' definition of a convention.[19]

Consequently, the explanatory power of Lewis' Signaling Games theory is restricted since it can only explain cases which accord with the limitations.

Both limitations point to an obvious argumentative move Lewis could have taken: he could have defined the conventional pattern in a different way using his "official" definition of a convention. According to my reconstruction of it (§4.1) the regularities can be *in action and attitude – i.e.*, involve any type of mental state for which one can have reasons, including beliefs, desires, and intentions – and the preferences may be quite different as long there are two strategy profiles which are such that there is a general

[18] For a recent empirical meta-study, see (Henrich et al. forthcoming).

[19] We could apply my *expected-utility* approach to allow for more deviations. But the real issue seems to me to be the communicative pattern that is wrongly described.

preference for general conformity to them.

Using the latter notion of a convention, the communicative pattern can be described in Gricean terms of speaker-meaning and understanding. First we interpret the states, the elements in the set T of a two-sided signaling problem, as types of mental states of the sender. Second, we interpret the receiver reactions, the elements in the set A, as types of mental states of the receiver. On this basis, we can define the sender strategies as contingency plans to send certain messages depending on her communicative intentions, desires, and beliefs. Likewise, we can define the receiver strategies as contingency plans to come to believe that the sender has a certain communicative intention when she signals a certain message, also depending on the receiver's beliefs and desires.

On this proposal, receivers only need to exhibit a uniform reaction with regard to linguistic understanding and not with regard to the "perlocutionary" effects. Hence, it's more plausible that there is a one-to-one mapping from states (to messages and from messages) to reactions. Moreover, by describing the communicative patterns in terms of communicative intentions and their recognition, the signal meanings that are determined by such signaling conventions are now proper literal meanings.

There is a further advantage of the proposal: As long as the senders' and receivers' preferences coincide enough with respect to message signaling and linguistic understanding, the ulterior goals the senders want to realize by signaling can be ones the receivers don't want to realize. Thereby the agents' preferences can be even more heterogeneous.

The last point can be clarified by distinguishing two kinds of preferences over the dependencies between states and receiver reactions.[20] *Near* preferences are preferences over the dependencies between states and a receiver's linguistic understanding of the respective message signaled by the sender. That is, preferences over $T \times A$ on the interpretation that the elements in T are types of mental states of a sender (*e.g.* defined in terms of Gricean in-

[20] This approach is preferable to the one of so-called "imperfect signaling systems". They are typically interpreted as being about far preferences and allow for disaligned preferences. According to them, the less aligned the preferences, the less information can be conveyed. In other words, the signals' meanings depend on how aligned the preferences are of the involved parties. These intuitions were mathematically studied in (Crawford and Sobel 1982) and (Spence 1973); van Rooy (2003:§6) illustrates their results for a finite case. I think this approach is more coherent for signaling systems with meaningful signals since it's best understood as a claim about how much information can be conveyed.

tentions) and that the elements in A are types of mental states of a receiver (*e.g.* defined in terms of beliefs of what the respective sender intended).[21]

Far preferences are preferences over the dependencies between states and the behaviors that receivers exhibit depending on their linguistic understanding of the signal given in the state. That is, preferences over $T \times A'$ where T is interpreted as above but A' is interpreted as Lewis originally proposed, *i.e.*, the elements in A' are *actions* of a receiver whose performances result from understanding the respective signal.

Using this distinction and Lewis' latter definition of a convention, we can assume that only the near preferences of the parties to a signaling convention need to coincide and this only to a limited extent: there only need to be two strategy profiles which are proper coordination equilibria. Importantly, two aspects are irrelevant for this to hold. First, it's irrelevant how the near preferences are over the *other* strategy profiles. Second, it plays little role how the far preferences are. But actually, this is not so clear since there seems to be a connection between learnability and robustness of a signaling pattern on the one hand and the relation between near and far preferences on the other hand: On condition that the signaling pattern should be learnable and be relatively robust, it seems that the far preferences constrain (but do not determine) the near preferences in some non-trivial way.

This argumentative move requires no change in the formal definition of a two-sided signaling problem but only a change in the interpretation of what the states and the receiver reactions are. Some coordination between senders and receivers is still required but only with regard to the near preferences. And it seems that we may assume this to be the case even in conditions of occasional conflicts of far interests. For it seems to be beneficial for a manipulating sender to have a conventional signaling system.

[21] Sperber and Wilson (1995:§1.11–1.12) draw a similar distinction between communicative and informative intentions. The latter are intentions of speakers to modify the cognitive environment of hearers, *e.g.* their beliefs; the former intentions are intentions of speakers that hearers recognize their informative intentions. In terms of this distinction, my proposal comes close to saying that what I call "near preferences" are preferences over sender states and the receivers' fulfilling a certain communicative intention depending on the signal sent.

5.3.2 Extension to other moods

Lewis considered only two moods (or speech act types): imperatives and indicatives. Can we generalize to other moods? There seems to be no problem to do so. It's made even easier on the basis of the proposal in the last section. According to the popular proposal of Bach and Harnish (1980), to perform a speech act of type μ is to express a (possibly complex) propositional attitude that is characteristic for μ with the intention that the receiver recognizes the attitude. A sincere, non-defective performance results in the receiver's recognizing the expressed attitude. On one understanding of "recognition," this amounts to coming to believe that the sender has the expressed attitude. I don't want to evaluate whether Bach and Harnish's proposal is convincing. But I want to show that their explications can be integrated in Lewis' theory. To illustrate the idea, I consider the example of *promises* and show what has to be done:

First, we need to assume that it is for S and R in their common interest to do their part in the respective two-sided signaling system.

Second, we have to assign a code number to the mood *Promise*. I use number 2.

Third, the set T of states and the set A of receiver reactions have to be defined suitably: T is defined as consisting of classes T_2 of (possibly complex) propositional attitudes (in particular: beliefs, desires, and intentions) which are characteristic for 2. Likewise, A is defined as consisting of classes A_2 of beliefs of the kind that the sender has this-and-that attitude, which are characteristic for mood 2. The sender and receiver strategies are as above. But we still have to extend the satisfaction-relation for the additional moods.

The characteristic attitude complex of *promises to receiver R to do some action b* consists in (i) the belief that the sender's (S) signal obligates her to do b, (ii) the intention to do b, and (iii) the intention that R believes that S's signal obligates S to do B and that S intends to do b (Bach and Harnish 1980:50). The receiver's characteristic attitude complex consists in the belief that S's signal obligates S to do b and that S intends to do b.

For each action b of a set B of actions whose performance senders can promise, T_2 has a suitable state, and likewise for A_2. Suitable states and reactions are ones that consist in the respective characteristic attitude complex from above.

Fourth, we need to single out suitable satisfaction conditions τ that can be attributed to promises. I suggest that the satisfaction condition τ of a promise to do b is *to do b* (with b being an action S can promise).

Fifth, we need to extend a general clause for the satisfaction conditions of promises:

(4) 2-satisfaction: If $\mu = 2$ (promise), then τ is 2-satisfied in situation s iff what almost every sender in s does (or not too much later) is τ.

5.3.3 Escaping the regress argument

In my short history about the conventionalist project in §1.1.4, I mentioned a popular regress argument with the conclusion that in general, conventions cannot determine the meaning of words (and other kinds of linguistic expressions) since this gets us in a regress: conventions already presuppose a meaningful language. The argument is a non-starter but it's instructive to see why. One of the classic proponents of the argument is Bertrand Russell who thinks that

> [i]t is natural to think of the meaning of a word as something conventional. This, however, is only true with great limitations. A new word can be added to an existing language by a mere convention, as is done, for instance, with new scientific terms. But the basis of a language is not conventional [. . .]. (Russell 1921:189)

His argument for the conclusion is that if we trace all the human languages back to its root language, then this language has meaningful words but it cannot be the case that the words have meanings in virtue of conventions:

> How these roots acquired their meanings is not known, but a conventional origin is clearly just as mythical as the social contract by which Hobbes and Rousseau supposed civil government to have been established. We can hardly suppose a parliament of hitherto speechless elders meeting together and agreeing to call a cow a cow and a wolf a wolf. (Russell 1921:190)

In the last sentence of the quote, Russell seems to justify his claim that expressions of the root language are not meaningful in virtue of conventions: He points out that it's implausible to assume that people convened to verbally agree on the expressions' meanings. We can state the justificatory premise as follows:

VA. Conventions are verbally performed agreements.

If Russell is right about this, then another language is required. But by assumption, there is no other language since it's already the root language. This rules out that the meanings of the words of this language are conventional. So, we have to assume that the root language consists of words that have meanings but *not* in virtue of conventions. So ultimately, the explanation why words are meaningful is not in terms of conventions. This interpretation is in line with how the passage ends:

> The association of words with their meanings must have grown up by some natural process, though at present the nature of the process is unknown. (Russell 1921:190)

In a different context, Quine (1976b) advanced a related argument: Mathematical sentences cannot be true in virtue of being derived from notational conventions since these conventions are again sentences. But these sentences are either themselves true in virtue of conventions or absolute truths. In the first case, we enter a regress which leads to the second case. In the second case, we do not analyze the truth of mathematical sentences in terms of conventions anymore.

However, having gone through Lewis' analysis of convention in the last chapter, we know that premise VA is wrong. For Lewis conventions can come into existence without requiring linguistic mediation. But if VA is false, the regress arguments don't get off the ground. Moreover, Lewis' Signaling Games theory explains in detail how signaling conventions determine signal meanings. Admittedly, this is not enough for linguistic meanings but I think it's more than enough to rebut the regress argument. Finally, the answer to the regress arguments makes it clear why a successful analysis of conventions as non-verbal agreements is so important for conventionalist use theories of meaning.

5.3.4 Is Lewis' Signaling Games account Gricean?

In his article *Meaning* (Grice 1957), Grice made an important distinction between two notions of *meaning,* of which he deemed only one to be relevant for the analysis of literal meaning. Is the notion of meaning Lewis uses in his theory faithful to Grice's proposal?

Grice called the irrelevant notion "natural meaning" and the relevant notion "non-natural meaning." Among the examples he gave, (5) is an example for natural meaning and (6) for non-natural meaning:

(5) Those spots mean (meant) measles.

(6) Those three rings on the bell (of the bus) mean that the "bus is full."

This kind of "non-natural meaning" is called "speaker-meaning." On the basis of the statement of Lewis' theory, it seems that it is unrelated to speaker-meaning.[22] For Lewis nowhere uses *speaker-meaning* or a cognate in the statement of his theory. But he claims that speaker-meaning "is a consequence of conventional signaling" (Lewis 2002:154). So, it seems that Lewis wanted there to be a connection between speaker-meaning and conventional signaling. More precisely, it seems that he wanted his account to be Gricean in the following sense:

SG-G. Normally, senders speaker-mean something (in *Grice*'s sense) when they utter a verbal expression conforming to a signaling convention they are party to.

On pp. 152–159 in *Convention* (Lewis 2002), Lewis provides a "proof" for this statement. However, contra Lewis, we need to make additional assumptions to derive the desired conclusion.[23]

Before I get into the details, I'd like to reflect on why Lewis wanted his account to be Gricean. Lewis does not tell us this. But we don't have to speculate a lot. For the truth of SG-G secures that in paradigmatic cases of signaling, the conditions of speaker-meaning are satisfied. So, these cases are about non-natural meaning and hence about the "right" kind of meaning for the analysis of literal meaning. (I'm not convinced by this. In the next chapter, I'll argue that a conventionalist account should not be Gricean in this sense.)

Now let us turn to the details of Lewis' proof. Lewis' strategy to establish SG-G consists of two steps. First, he assumes that CS is the case:

CS. There is a normal situation in which a sender of a conventional signaling system produces ("utters") a signal in a way conforming to the signaling convention.

Then, Lewis shows that in a conventional signaling situation CS what the sender did satisfies the analysis of speaker-meaning. Recall that Grice's analysis of speaker-meaning (in a canonical version offered in Grice 1969) is

[22] On the proposal I offer in §5.3.1, the theory is Gricean for I restate the conventional patterns in terms of speaker-meaning and understanding.

[23] While I do not argue for it here, this objection is *mutatis mutandis* also valid for Bennett (1976:179–181) whose allegedly simpler "proof" has more problems.

as follows:[24]

> "U meant something by uttering x" is true iff, for some audience A, U uttered x intending
>
> 1. A to produce a particular response r
> 2. A to think (recognize) that U intends 1.
> 3. A to fulfill 1. on the basis of his fulfillment of 2. (Grice 1969:151)

Lewis thought that he could establish clauses 1, 2, and 3 by examining the practical reasoning that justifies the sender to produce a certain signal (Lewis 2002:155). I object that plausibly, not all of the clauses are satisfied if CS is assumed to be true.

First, Lewis does not use Grice's analysis but his personal restatement in which the clauses above are replaced by the following ones (where U = I, the sender; A = you, the receiver; and utter x = signal m):[25]

LSM1. I signal m with the intention that you do a.

LSM2. I expect you to recognize my intention that you do a, when you observe that I signal m.

LSM3. I expect your recognition of my intention to be effective in bringing it about that you do a. I do not regard it as a foregone conclusion that my action will bring it about that you do a, whether or not you recognize my intention that you do a.

It is unfortunate that Lewis used his personal restatement. For a clause-by-clause comparison of the two analyses reveals some non-trivial differences. This leads to the worry that even if LSM1–3 are satisfied, Grice's clauses aren't. To establish this claim, we would need to find out whether the corresponding clauses are equivalent or if not, which semantic relations hold between them. Since there is other criticism, I spare me this job.

Second, Lewis understands "(to) intend" as "(to) expect and want" (p. 153) and it is clear from his proof on p. 155 that he uses "(to) want"

[24] This is more or less Grice's original analysis as of (Grice 1957) which is known to have defects. I discuss a later analysis in §6.2.4.

[25] This is a reformulation of (Lewis 2002:154 ff.) to harmonize it with the way I presented Lewis' theory. I changed "intention to produce r" to "intention that you do a", "I do [signal] σ" to "I signal m", "my action will produce r" to "my action will bring it about that you do a", and "producing your response" to "bringing it about that you do a." Nothing hinges on the changes.

synonymously with "(to) desire." So, on Lewis' proposal, intentions are not attitudes *sui generis* but reduced to beliefs (or expectations) and desires. But this explication of an intention is quite peculiar. For it seems possible that someone can expect and desire something without intending it. Also, it's not obvious that it is the same notion as the one Grice uses which has its roots in folk-psychology which assumes at least three different central attitudes, namely beliefs, desires, *and intentions*. It's also not the standard notion in the BDI-literature; in particular, it's not the same as the popular definition of "intention" along the lines of "choice with commitment" which has been proposed by Cohen and Levesque (1990). Thus, Lewis' explication goes against our pre-theoretic understanding, against folk-psychology, and against the established understanding in the BDI-literature. So, at the very least, Lewis' explication of intention requires a justification but he didn't offer any.[26]

Third, it seems that LSM2 can only be established if we make substantial additional assumptions. Lewis reasons on p. 155 as follows:[27]

> I expect you to infer [t] upon observing that I [signal m]. I expect you to recognize my desire that you do a, conditionally upon [t]. I expect you to recognize my expectation that I can [bring it about that you do a by signaling m]. So, I expect you to recognize my intention [that you do a], when you observe that I [signal m]. (Lewis 2002:155)

While I see that the sentences before the conclusion "So, ..." follow from CS, I don't see how the conclusion follows from the premises he uses (*i.e.* the sentences before the "So, ..."). So, I provide my own reconstruction. To establish LSM2, let me first state CS more precisely as CS′:

CS′. There is a normal situation in which a sender of a conventional signaling system produces ("utters") a signal in a way conforming to the signaling convention iff

a. There is a regularity R in situations which are two-sided signaling problems G among you and me, and R is a conventional signaling system $\langle \sigma, \rho \rangle$ among us.

b. Let m be a message, t a state the world can be in, and a a response such that $m = \sigma(t)$ and $a = \rho(m)$.

[26] But see (Kemmerling 1979) who suggests a reconstruction of speaker-meaning in terms of beliefs and desires (without intentions).

[27] The brackets indicate some changes I made to harmonize the quote with my presentation.

c. There is a situation t which is a two-sided signaling problem of type G where I am the sender and you are the receiver.
d. I have observed that t holds.
e. I signal m in conformity to our convention.
f. The convention is "perfect," that is, there are no exceptions to it.

We also need to agree on what "to recognize (an intention)" means since Lewis is silent about it. We can build on two observations. First, Grice used "to think" (to believe) and "to recognize" synonymously in (Grice 1969:151). This justifies the following conditional *for the purposes at hand*:[28]

A1. If x believes y, then x recognizes y.

The second observation is that Lewis generates the hierarchy of mutual expectations that the parties to a convention conform to it from common knowledge and *expectations* seem here to be the same as *beliefs*.[29] This justifies us to assume:

A1′. If x expects y, then x recognizes y.

Lewis defines "expect" along the following lines:[30]

A2. x expects y iff x has reason to believe y and x is rational to the (high enough) degree n.

So, to expect something is to have an actual belief, acquired by having a reason to believe and thinking about it for long enough. Note that on the basis of A2, it follows from the meaning of "x expects y" that "x believes y." So A1′ entails A1 (and *vice versa*; here the first direction is relevant).

Now, assuming A1′, we can establish LSM2 by establishing LSM2′:

LSM2′. I expect that (you expect that (I intend that you do a), when you observe that I signal m).

To establish LSM2′, we need the following three rationality assumptions:

A3. I reason rationally to a high enough degree.

[28] In the context of Grice's recognition of intentions, I think understanding recognition as *believing* or *coming to believe* is tenable; in other contexts, I think the claim is questionable.
[29] See p. 52 in (Lewis 2002).
[30] This interpretation is supported by Lewis' use of common knowledge in his definition of conventions and his definition of common knowledge on pp. 52–60 in (Lewis 2002).

A4. I expect that you expect that I am rational.

A5. I expect that you are rational.

Remember that Lewis' definition of a convention does not require the agents to be rational or say anything about their mutual rationality assumptions. Though it seems that in the present context, he takes A3, A4, and A5 for granted.[31]

From assumption CS′, A3–A5 it follows that:

(7) You observe that I signal m, and that

(8) I expect that you expect that I will signal m if and only if t holds.

From (7), (8), and A4 it follows that:

LSM2′a. I expect that (you expect that t holds, when you observe that I signal m).

From assumption CS′ and the definition of a two-sided signaling problem it follows that I desire that you do a, conditionally upon t. From this and the common knowledge assumption of the definition of a convention it follows that

LSM2′b. I expect that you expect that I desire that you do a, conditionally upon t.

From LSM2′a and LSM2′b, and the rationality premises A3 and A4 it follows together that

LSM2′c. I expect that you expect that (I desire that you do a, when you observe that I signal m).

From CS′ it follows that I expect you to conform to R. When you observe m, you conform to R by doing a. So, I expect that you do a when you observe that I signal m. From this and the common knowledge assumption of the definition of a convention it follows that

LSM2′d. I expect that you expect that I expect that (you do a, when you observe that I signal m).

[31] This also seems to be the received view. For example, Savigny (1985:87) is convinced that Lewis' presentation leaves "no room for doubt that [...] all but 'children and the feeble-minded' are sufficiently rational."

Having established LSM2′c, and LSM2′d we can now establish LSM2′. Recall that to intend something is to expect and desire it. Thereby, LSM2′d entails the "expect"-part of LSM2′. LSM2′c entails the "desire"-part. Thereby LSM2′ has been established.

Reflecting on these tricky steps, there are three important points. (i) To establish LSM2, we had to assume A1–5. In particular, we had to assume that the sender is to some degree reasoning rationally (A3), that the sender believes that the receiver believes that she reasons to some degree rationally (A4), that the sender believes that the receiver reasons to some degree rationally (A5). (ii) These "rationality"-assumptions are by no means trivial. But if Lewis' theory should be Gricean as he presumably wanted, we can make it so by imposing assumptions A1–5. (iii) Since the reasoning is quite unprincipled and needed substantial additional assumptions, I think Lewis was too quick in calling it a "proof."

5.4 Evaluation

Lewis' theory is simple, conceptually clear, highly informative, and modular. But his account fails to meet the bar of an adequate account. I consider first some criticism and then turn to the adequacy of the theory.

5.4.1 Criticism

Lewis (2002:160 ff.) himself was probably the fiercest critic of his theory. On the basis of the changes I've suggested in §5.3, his verdict could have been milder. Yet, some of his points remain, among them being: (i) Signaling languages consist of a closed and finite set of sentences. There is no room for creative language use, *i.e.* the use of new sentences built up from parts of old sentences. (ii) The receiver is in a position to and has an interest in making true any imperative sentence. This is a rather limited understanding of the role of the audience.

To Lewis' criticism, I add the following further limitations:

First, signals in a signaling language have no structure. But linguistic signals are structured, composed of mood indicators and words in specific syntactic arrangements. If signals are unstructured, one cannot explain on the basis of the theory how moods are indicated, and how syntactic and semantic relations between expressions are constituted. For example there is nothing in Lewis' Signaling Games theory which allows to explain how the truth conditions of an indicative of the form "A and B" depends on its parts. I think, however, that there is a rejoinder to this objection. I'll return to it in §7.2.3 in the context of Evolutionary Signaling Games.

Second, since Lewis' Signaling Games theory is not stated in terms of normative notions with a demanding character, semantic normativity cannot be explained. In particular, his notion of a convention does not have a demanding character. Hence, meaning is not normative in the required sense (see chapter 2). But an explanation of semantic normativity is required for an adequate conventionalist account (§1.4).

Third, Lewis' theory in its original statement is in fact not Gricean (on my proposal in §5.3.1 it is). This is undesirable, if Grice is right. For it is then an open question whether we are justified to say that signals of a signaling convention have a *literal meaning*. And Lewis presumably wanted to make his theory Gricean. By making the rationality-assumptions stated in §5.3.4, we can make it Gricean. But thereby, the account has two problematic consequences.

On the one hand, the account then entails a rationalistic agent conception. According to it, agents must be Intellectual (in the sense of being Bayesians who consciously deliberate). Or they must be habitually acting rational agents whose actions are in fact rational while they often don't deliberate consciously but behave habitually.

To me it seems that such a rationalistic commitment is not desirable. Humans act sometimes irrationally. They make mistakes and their habits are often not corrected even if they learn that the actions they perform are irrational. So, it's not plausible to assume that agents act always or almost always rational. It is much more plausible to assume that the agents' underlying cognitive processes tend to bring about reasoning and acting that has adapted from the historical environment of their predecessors. In certain circumstances this leads to agents reasoning and acting rationally. If so, this shifts the study to the cognitive processes and the circumstances.

On the other hand, by being Gricean, Lewis' account has to cope with the problems related to the use of Gricean speaker-meaning. I won't go into this debate here, but will discuss it in the next chapter. But I think that it is fair to say that there is no convincing band aid for Grice's analysis of speaker-meaning after 50 years of counterexamples and remedies. So, being Gricean is undesirable.

So, a conventionalist account which can do without rationality assumptions is preferable to one which has to make such assumptions, *e.g.* Lewis' Signaling Games theory (and also his Actual Language Relation theory, for that matter). Doing without rationality assumptions should, of course, not be understood in a way that rules out that agents often act and reason rationally. But it should be explained how it comes that they act and reason rationally.

Moreover, a better conventionalist account should not be Gricean. This is also not meant to rule out that sometimes, senders satisfy the conditions of Gricean speaker-meaning when they use language. But it should not be the case that it is almost always the case.

5.4.2 Adequacy of Lewis' Signaling Games theory

To complete the evaluation, let us consider the adequacy of Lewis' theory according to the adequacy condition I proposed in §1.4. The condition consists of three sets of desiderata: A first relating to an account of conventions, a second to an account of social norms, and a third to an account of meaning.

Conventions We know the result with regard to the account of conventions. By the way Lewis' theory is designed, it depends on an independent account of conventions. Lewis' used his account as developed in *Convention.* Thereby, his Signaling Games theory inherits the strengths and weaknesses of his account of conventions. For example, (i) dispositional conventions cannot be explained by it (DesC1) and (ii) the account tends wrongly towards a rationalistic interpretation according to which agents act the way they act because they deliberate (DesC5). The last problem is more pressing for signaling conventions. Here, the rationalistic interpretation seems unavoidable if Lewis insists that his theory should be Gricean.

Social norms With regard to the account of norms, there is not much to say. Since his notion of a convention is not demandingly normative and the rest of his Signaling Games theory does not contain demandingly normative statements, there is nothing to explain semantic normativity.

Meaning Lewis' theory of meaning is not adequate, partly because of the lack of an account of social norms. For this reason, Lewis' account does not satisfy the requirements related to the normative features of meaning (DesM1). Relatedly, there is no meaning in virtue of social norms (DesM2). Lewis' theory also does not seem to allow for a plausible conception of a linguistic mistake (DesM3) or a plausible conception of a public language (DesM4).

The issue with linguistic mistakes is the following. If we only use the vocabulary of Lewis' theory, then to make a linguistic mistake amounts to deviate from a signaling convention. Since signaling conventions have a recommending but not a demanding character, mistakes in this sense are violations of instrumental rationality. Someone making such a mistake violates a prudential *ought*. But I've argued in §2.2.3 that to make a *semantic* mistake (which is the relevant kind of linguistic mistakes for our purposes here) is not to violate a prudential *ought* but to violate an *ought* with a demanding character. Lewis' theory does not allow us to explain semantic mistakes in this sense since there is no normative notion in his theory that has the required demanding character.

Another inadequacy of Lewis' theory is that it does not provide a plausible solution to Humpty Dumpty's problem. For almost all members of a population are needed to determine a signal's meaning. However, this seems to be wrong (DesM5). Lewis' theory can explain some central meaning facts (*e.g.* how physical items can have a meaning). But the theory fails to explain linguistic structure and thereby linguistic meaning (DesM6).

5.5 Summary

As Lewis observed, his theory in *Convention* was that it is just a theory of a fragment of actual language use (see *e.g.* Lewis 2002:143, 161). But while Lewis' theory as it stands is not adequate, we should not give up what's appealing about it: Central about language use are the particular interactions which require coordination. Conventions relate a sender's behavior to a receiver's reaction. And conventions are what make coordinations successful and lead to stable use patterns which seem to have all the features required to say that the expressions used have a meaning.

Looking back, I'd like to highlight the following points. First, Lewis' two definitions actually have an effect on Lewis' Signaling Games theory. If he had used his later "official" definition, he could have avoided the implausible assumption that receiver reactions consist in actions. This led us to introduce the notion of a near preference which does not relate states and receiver reactions but states and receiver understandings. By restating the communicative patterns in terms of speaker-meaning and understanding, the meanings Lewis' theory determines became literal meanings. Second, Lewis' theory can be generalized to cover further moods. Third, Lewis' theory is either not Gricean or Gricean and rationalistic. Fourth, the regress arguments of Russell and Quine can be escaped by using Lewis' theory.

Chapter 6

Actual Language Relations

> There is some relation R – call it the *actual-language relation* – such that a language $\mathcal{L}$ is a language of a population P iff $R(\mathcal{L}, P)$, and the problem – call it the *actual-language-relation problem* – is to say what R is.
>
> *Actual-language relations*
> Stephen Schiffer

In this chapter, we turn to the second paradigm of conventionalist accounts: "Actual Language Relation" accounts. Such accounts employ a so-called "actual-language" relation to answer the question in virtue of what expressions mean what they do: a sentence's meaning is determined by *the language the population uses.* The task is then to say what it is for a population to use a language. David Lewis provided the standard account on which I will focus. His proposal was to conceive of meanings as abstract objects, with a preference to use an intensional semantics, *e.g.* propositions for sentences. Natural languages are in a sense functions from sentences to meanings. Let us call such functions "abstract languages." Lewis suggested that a population P uses an abstract language $\mathcal{L}$ iff there is a convention of truthfulness and trust in $\mathcal{L}$ among members of P, which is sustained by an interest in communication. For a speaker of P to be truthful in $\mathcal{L}$ is to try to avoid uttering sentences not true in $\mathcal{L}$. For a hearer of P to be trusting in $\mathcal{L}$ is to tend to believe that sentences uttered by speakers of P are true in $\mathcal{L}$. Thus, the basic tenets of Lewis' theory can be stated in three theses:

ALR1. The meaning of an expression in a natural language of a population is determined by the population's use of an abstract language corresponding to the natural language.

ALR2. A population uses an abstract language $\mathcal{L}$ iff there are conventions of truthfulness and trust in $\mathcal{L}$ among the members of the population.

ALR3. There are conventions of truthfulness and trust in an abstract language $\mathcal{L}$ among members of a population P iff speakers of P try to avoid uttering sentences not true in $\mathcal{L}$ (truthfulness) and hearers of P tend to believe that sentences uttered by speakers of P are true in $\mathcal{L}$ (trust).

The plan for this chapter is as follows. Lewis' elementary theory is developed and extended in §6.1, with a special interest in a theoretically important objection called the "meaning without use" problem which leads to a modification of the theory. The resulting theory and other related accounts are evaluated in §6.2. As usual, the chapter ends with a summary in §6.3.

6.1 The elementary theory and its extensions

Lewis (2002) first presented his Actual Language Relation theory in chapter V of his book *Convention.* I base my presentation on his later revision in *Languages and language* (Lewis 1975), focusing for the moment on assertive uses of indicatives.[1] The theses ALR1–3 characterizing Lewis' theory indicate its ingredients: (i) an actual-language relation defined in terms of (ii) conventions of a particular sort, namely ones having (iii) regularities of truthfulness and trust in (iv) a particular abstract language among (v) members of some population. Let us have a closer look at them with the goal of stating a general meaning determination claim.

6.1.1 Regularities of truthfulness and trust

Lewis defines the actual-language relation in terms of *conventions of truthfulness and trust.* Conventions of truthfulness and trust are a certain kind of conventions in his sense (according to his later definition in *Languages and language,* see chapter 4). What is special about them is that their regularity is one of being truthful and trusting in an abstract language. Lewis'

[1] The main versions or clarifications of his account are: (Lewis 2002, 1975, 1976, 1992). The 1975 version was prompted by criticism and suggestions by Bennett (1973, 1976) and Schiffer (1993). A perspicuous presentation and discussion of Lewis' account in *Convention* can be found in (Kemmerling 1976:74–128). Helpful presentations of Lewis' theory are in (Schiffer 1993), (Schwarz 2009:§10.2), and (Weatherson 2009:§2.2 ff.).

formulation of the regularity has several noteworthy features: First, the regularities are stated metalinguistically. Instead of stating them directly in terms of attitudes with a certain content (which is not necessarily about the truth of sentences), they are stated in terms of beliefs about the truth of sentences. *E.g.* Jones is truthful in English when he utters "It's raining" if he believes that the *sentence* "it's raining" is true in English (and not, more mundanely, if he *believes that it is raining*) – and likewise for being trustful. Certainly, it's not necessary to have such metalinguistic beliefs in order to be able to use a language. Lewis does not elaborate on why he choose to formulate it this way and it's unclear to me why he did so. Instead of understanding Lewis' proposal as requiring the language users to have metalinguistic beliefs, I think that we can faithfully interpret it as characterizing types of mental states that are not necessarily metalinguistic by making the additional assumption that one believes that a sentence is true in an abstract language if one believes (/wants/...) what it expresses in the language. Truthfulness then amounts to "Speaker S doesn't utter sentence s unless she believes (/wants/...) that p (and p is the proposition expressed by s)."

A second feature of the regularity is that the overall analysis is neither speaker- nor hearer-centered. Both are assigned a role. Since being truthful is common knowledge, hearers can expect speakers to be truthful. This seems to amount to trust and raises the question why we can't simplify the regularity to one of truthfulness. There are two problems in this area: the truthfulness-by-silence problem and the meaning-without-use problem. I turn to the former here and to the latter below in §6.1.5.

Lewis' account in *Convention* had only regularities of truthfulness. Stephen Schiffer observed that this is problematic (*cf.* Lewis 1975:398 ff.): Lewis' old account had the contra-intuitive consequence that it was not able to distinguish between a population using a language and its using other languages that are identical but include a class of garbage sentences that speakers would never utter. Since these sentences would never be uttered, they would never be uttered *falsely.* Hence, the truthfulness-condition is satisfied vacuously since the condition to speak truthfully is conditional on uttering a sentence and the sentences in question are never uttered. The new formulation including a regularity of trust avoids the truthfulness-by-silence problem. For expectation of truthfulness (following from common knowledge of the convention) is not yet trust. In (Lewis 1992:107), Lewis made this more precise: An agent A is trusting in an abstract language $\mathcal{L}$ iff for all sentences $s \in \mathcal{L}$, A's subjective likelihood ratio

$$\frac{p(s \text{ will be uttered} \mid s \text{ is true in } \mathcal{L})}{p(s \text{ will be uttered} \mid s \text{ is false in } \mathcal{L})}$$

which measures the extent to which the truth of s in $\mathcal{L}$ is confirmed when s is uttered, exceeds 1 by enough margin. This formulation allows for minute distinctions in the subjective probabilities of a hearer by comparing conditional probabilities. This condition is not satisfied by mere expectation of truthfulness in $\mathcal{L}$ since trust requires that the posterior degree of belief that an uttered sentence is false *is low* but this is not entailed by an expectation of truthfulness. This is one reason for the regularity of trust.

The third notable feature has to do with coordination. The old regularities of truthfulness resulted in one-sided coordination problems. Speakers were assumed to coordinate very indirectly with past and future speakers – and not with the present hearer. Postulating such a coordination feels artificial and is quite different from the two-sided coordination problems used in Lewis' Signaling Games theory between the particular speakers and hearers in instances of communication.

The new regularities restore proper coordination: speaker and hearer are assumed to have a common interest in the hearer's acquiring whatever belief the speaker wants her to acquire by means of using and understanding sentences. The speaker's problem is that there are different sentences she can utter which could lead to the desired result. The best thing she can do depends on how the hearer reacts to the utterance. The hearer is in a similar predicament. She wants to acquire the belief the speaker wants her to acquire. The hearer's problem is to react to an uttered sentence in a suitable way. But she could come to acquire different beliefs and the best thing she could do depends on the speaker's choice. So, they face a coordination problem of the familiar sort. This way of stating the conventional regularity became possible by Lewis' revision of his notion of a convention, now allowing not only for regularities in action (as before) but also in *action and beliefs* (§4.1).[2]

[2] Jonathan Bennett has improved on Lewis' definition and Lewis (1976:386) accepted the proposal, *cf.* (Bennett 1973:§5) and (Bennett 1976:§55).

6.1.2 The actual-language relation

Having defined conventions of truthfulness and trust in an abstract language, the definition of the actual-language relation is obvious: For populations P and abstract languages $\mathcal{L}$: P uses $\mathcal{L}$ iff there is a convention of truthfulness and trust in $\mathcal{L}$ among members of P, sustained by an interest in communication. Importantly, the definition allows that a population uses more than one language in this sense at a time. The definition looks less innocent when we unpack the definition of a convention (*cf.* Lewis 1975:384 ff.):

There is a convention of truthfulness and trust in $\mathcal{L}$ among members of P iff

1. Everyone conforms to a regularity of truthfulness and trust in $\mathcal{L}$.
2. Everyone believes that the others conform to the regularity.
3. This belief that the others conform to the regularity gives everyone a good and decisive reason to conform to the regularity herself.
4. There is in P a general preference for general conformity to the regularity of truthfulness and trust in $\mathcal{L}$ rather than slightly-less-than-general conformity.
5. The regularity of truthfulness and trust in $\mathcal{L}$ (R) has [at least one alternative R', namely truthfulness and trust in another language $\mathcal{L}'$, such that the belief that the others conformed to R' would give everyone a good and decisive practical or epistemic reason to conform to R' likewise; such that there is a general preference for general conformity to R' rather than slightly-less-than-general conformity to R'; and such that there is normally no way of conforming to R and R' both.]
6. All these facts are common knowledge in P.

I present clause 5 in a more explicit than Lewis since it's a clause we'll return to below. According to it, an alternative to a regularity of truthfulness and trust in an abstract language $\mathcal{L}$ is a regularity in another language $\mathcal{L}'$ that "does not overlap $\mathcal{L}$ in such a way that it is possible to be truthful and trusting simultaneously in $\mathcal{L}$ and in $\mathcal{L}'$" (p. 385). Importantly, not all regularities of truthfulness and trust in other languages have to be alternatives in this sense.[3]

Lewis thought that his analysis was justified by the ordinary use of "using a language." But both the requirement of common knowledge of the alternatives and of the regularity's being sustained by a common interest in communication have been criticized by Burge (1975:250 ff.) and Kemmerling (1976:117–119): (i) A population can use a language without knowing

[3] I thank Max Kölbel and Wolfgang Schwarz for a helpful discussion on the issue.

about its alternatives (required by clause 6). (ii) Suppose members of a population desire to speak the gods' language. They believe that the gods don't switch languages. Then there is no alternative to the language the members desire to use (required by clause 5). (iii) Suppose that a population speaks a language against its interest because its members are threatened by its usurpers. Then there is no common interest in communication in the language (required by the definition of the actual-language relation). While I think that these counterexamples have a point, I don't think that they seriously threaten Lewis' theory. One could simply say that common knowledge of alternatives is not necessary and drop the "sustained by an interest in communication" clause.[4]

6.1.3 Meaning determination claims

Since abstract languages determine for each sentence a meaning and the actual-language relation pairs populations with such a language, we can now state how conventions determine a sentence's meaning:

ALR-S. Sentence s means m in $\mathcal{L}$ used by P iff P uses $\mathcal{L}$ and $\mathcal{L}(s) = m$.

The actual-language relation does not allow us to determine the meaning of words (and sub-sentential expressions in general). To allow for this, a second relation between populations and grammars is introduced (*cf.* Lewis 1975:389–391). Let us call it the "actual-grammar relation." Grammars are assumed here to be *semantically interpreted* grammars generating not only well-formed expressions according to a set of operations but expressions paired with suitable meanings.[5] Let us designate a grammar that generates an abstract language $\mathcal{L}$ by "$\Gamma_{\mathcal{L}}$" and the abstract language generated by a grammar Γ by "$\mathcal{L}_{\Gamma}$". Lewis suggested that a grammar $\Gamma_{\mathcal{L}}$ of an actually used language $\mathcal{L}$ determines the meanings of *all* expressions – sub-sentential or not.[6]

ALR-E. Expression e means m in $\mathcal{L}_{\Gamma}$ of P iff P uses Γ and Γ pairs e with m.

[4] Schwarz (2009:196) assesses the scenarios similarly. The moves considered here have already been suggested by Kemmerling.

[5] The notion of "generation" is one from the theory of formal languages, defined in terms of the expressions that are derivable from a grammar; see (Hopcroft et al. 1979).

[6] *E.g.* by means of a categorial grammar as Lewis proposed in (Lewis 1970).

The step from ALR-S to ALR-E is not trivial. For there are many weakly equivalent grammars generating the same language but possibly assigning semantic values of wildly different types to the same sub-sentential expression. Thus, there is an element of choice. Lewis was aware of this indeterminacy of his proposal and proposed that we should be indifferent between different grammars on condition that they generate the same abstract language (that is to say that the grammars must generate the same set of sentences pairing them with the same meanings). An alternative to this proposal would be to select grammars in an objective way by giving them a cognitive role, say in terms of mental realization. Lewis disliked this alternative since one would otherwise engage "in risky speculation about the secret workings of the brain" (Lewis 1992:110 fn. 6). For this reason, the meanings of sub-sentential expressions are only partially determined. (I emphasize the distinction between ALR-S and ALR-E to prepare for the discussion of the meaning-without-use problem in §6.1.5.)

6.1.4 Extensions to complexer languages

The abstract languages populations can use are highly restricted since there is no provision for indexical, ambiguous, or non-indicative sentences (Lewis 1975:387 ff.). Lewis offers extensions to his theory in various ways. Not all of them are convincing. For example, he defines "ambiguous languages" $\mathcal{L}$ as functions from sentences to finite sets of alternative meanings and restates the regularity of truthfulness and trust in an ambiguous language $\mathcal{L}$ as follows:[7]

> We can say that a sentence s is true in $\mathcal{L}$ at w under some meaning if and only if w belongs to some member of $\mathcal{L}(s)$. We can say that a sentence s is true in $\mathcal{L}$ under some meaning if and only if our actual world belongs to some member of $\mathcal{L}(s)$. We can say that someone is (minimally) truthful in $\mathcal{L}$ if he tries not to utter any sentence s of $\mathcal{L}$ unless s is true in $\mathcal{L}$ under some meaning. He is trusting if he believes an uttered sentence of $\mathcal{L}$ to be true in $\mathcal{L}$ under some meaning.
>
> (Lewis 1975:387)

This way of stating the regularity also seems to face a truthfulness-by-silence problem. For on the basis of this regularity, Lewis' theory cannot distinguish between a population using a language and its using languages

[7] Lewis uses "σ" where I use "s".

that are identical but assign the sentences some extra meanings such that the sentences would never be used in accordance with these meanings. Obviously, one would have to strengthen the truthfulness- and trust-conditions along the following lines: For all meanings m of a sentence s in $\mathcal{L}$: There are occasions where one would utter and understand s in accordance with m and someone wouldn't utter s in accordance with m, unless she believed that the actual world is part of m (understanding propositions as sets).

The extension to indexicals proceeds similarly (also p. 387). First, a class of *indexical languages* $\mathcal{L}$ is defined which are mappings from sentences to functions from occasions o of utterance to truth conditions. Second, the truth-in-$\mathcal{L}$ predicate is relativized to such occasions o. Third, the regularity is restated.[8]

The arguably most interesting extension is the one to *polymodal languages* $\mathcal{L}$ that are functions from sentences to pairs of code numbers and truth conditions (p. 387 f.). The code numbers have the same role as in his Signaling Games theory (§5.2), they indicate the mood μ of a sentence (like indicative, imperative, ...). *E.g.* the sentence "It's cold in here" can be used to describe how the temperature is in a classroom and it can be used to command someone to bring it about that it is not cold anymore. To this end, we could assign the sentence the following complex meanings $\langle 0, \textit{that it is cold in here}\rangle$ and $\langle 1, \textit{that it not be cold in here}\rangle$, together with a redefinition of being truthful and trusting as follows. A sentence is true in $\mathcal{L}$ iff the actual world is an element of the truth condition (understanding truth conditions as sets of possible worlds). Again, the regularities are restated, now parametrized to each mood:

- **Indicatives**: Someone is truthful in $\mathcal{L}$ with respect to indicatives if she tries not to utter any indicative sentence of $\mathcal{L}$ which is not true in $\mathcal{L}$. Someone is trusting with respect to indicatives if she believes uttered indicatives sentences to be true in $\mathcal{L}$.
- **Imperatives**: Someone is is truthful in $\mathcal{L}$ with respect to imperatives if she tries to act in such a way as to make true in $\mathcal{L}$ any imperative sentence of $\mathcal{L}$ that is addressed to her by someone in a relation of authority to her. Someone is trusting with respect to imperatives if she expects her utterance

[8] Lewis (1975:388) accounts for the intuitions of semantic externalists similarly, namely by introducing a "causal history entry" which truth conditions in the range of the function depend upon. Jackman (1998:300–302) rejects this approach in my opinion unconvincingly by pointing out that thereby Lewis models externalistic dependencies on the model of indexicals but thereby these two distinct phenomena become false friends.

of an imperative sentence in $\mathcal{L}$ to result in the addressee's acting in such a way so as to make that sentence true in $\mathcal{L}$, provided that she is in a relation of authority to the addressee.

- . . . and so on for other moods.

This way of defining what it is to be truthful and trusting relative to a code number amounts to an explication of moods or speech act types. Given a set of such mood explications, Lewis' theory is applicable to a wide range of language uses. Still, this is not satisfying. Kemmerling (1976:103–106) objected that the mood explications are implausible. And indeed, without an account of social norms, the issue cannot be fully addressed, as I've argued in §2.5 using the example of *to inform someone* and *to promise something.*

The extension to polymodal languages has some noteworthy consequences. First, it's not guaranteed anymore that a sentence in one mood has the same truth condition as in another mood since the truth condition for a sentence is defined independently for each mood. My example of "It's cold in here" illustrates this point. Consequently, it's also not guaranteed that any expression means the same in all contexts of its use. That is, Lewis' theory allows that "apple" means *apple* in indicatives and *non-apple* in imperatives. The importance of the point can be downplayed. For such a situation would hardly ever obtain since it's implausible that a convention of truthfulness and trust in such a language would obtain. But the implausibility is not explained by the theory.

Second, the extension has consequences for the kind of grammar used to determine what sub-sentential expressions mean. There is a complication from going from a complex meaning ⟨*code number*, *truth condition*⟩ to a *truth-condition* meaning. Where in the grammar is the code number introduced? If a sentence has several complex meanings – say one for each mood – whose truth-condition components vary, what is the truth-conditional meaning for the sentence? And likewise, how can the truth-conditional meanings for sub-sentential expressions be derived in such scenarios? These questions receive no obvious answers.[9] I think that one has to add the following claims to Lewis' theory: (i) All admissible polymodal languages assign to each sentence only complex meanings that have the same truth-condition component for each mood. (ii) The admissible grammars for ad-

[9] Kemmerling (1976:101 ff.) seems to criticize Lewis' former theory in *Convention* in a similar way.

missible polymodal languages introduce the code numbers only at the level of sentences.

6.1.5 Meaning without use

Stephen Schiffer (1993, 2006) and John O'Leary-Hawthorne (1990, 1993) observed that Lewis' theory has the meaning-without-use problem. It is a problem not just for Lewis' but for use theories of meaning in general. It results from the way use theories explain an expression's meaning, namely in terms of its use.

But it seems that at *any time*, there are actually unused expressions, for example very long and/or complicated expressions (involving embeddings, garden pathing, ...), which are meaningful even *before* they are actually used and (at least some of them) wouldn't readily be understood by typical members of the relevant community. Such expressions seemingly follow the English grammar. For this reason, many are inclined to count them as belonging to English (but this is contested, see below). They typically contain *parts* that complicate use and understanding, like: "the former", "the latter", "the third last counted from the middle", "the friend who saw the relative who is the father of the twins who direct the studio that is yellow died", *etc.*

The problem is that the meaning of actually unused expressions cannot be explained in terms of their actual use. Thus, it seems that use theories of meaning have an explanatory gap.[10]

The problem for Lewis' theory The specific problem for Lewis' theory (in the version discussed here) is as follows.[11] Very long sentences too complex for one to utter and understand them satisfy the truthfulness- and trust-conditions. Since there is no use at all, truthfulness goes through trivially. Trust is also satisfied. Recall that the condition is stated in terms

[10] To be clear, the issue is *not* neologisms that have so far no use, or gibberish that is used but meaningless or otherwise semantically defective.

[11] *Cf.* (Lewis 1992) which is a reaction to (O'Leary-Hawthorne 1990). Lewis rejects Hawthorne's objection but accepts his conclusion which I discuss in the main text. Hawthorne argues that Lewis' trust-condition is undefined since the unconditional utterance probabilities are 0 for very long sentences. Lewis' thinks that we should assume that they never get 0 but only very small. Other early discussions of the problem are in (Platts 1997:88–92), (Chomsky 1980:83–85), and (Blackburn 1984:127–130).

of conditional subjective probabilities. For very long sentences, most of the subjective probability mass would go to hypotheses which are unrelated to the truth and falsity of the respective sentences. Arguably, the remaining minute subjective probability would go in equal shares to their truth and falsity – and thus in effect to different abstract languages.

The result is that Lewis' actual-language relation cannot distinguish between a population using one of the candidate languages or another, even if they assign different truth conditions to the unuttered sentences. Notice that the problem for Lewis' theory pertains to expressions that (i) are not actually used and (ii) would not be understood, even after some reflection. The second condition is important. For actually unuttered expressions might still be understood. So, for an expression to be so far unuttered is alone not problematic since a stable disposition to understand utterances of it in a specific way would allow us to distinguish between different candidate languages.

"Solutions" There are different reactions to this problem. Probably there is no "straight" solution since people have conflicting intuitions. Loar (1976:158) denies that there is a problem in the first place since these "'sentences' are not really part of our language at all – after all, we cannot understand them!" His proposal is to restrict the class of abstract languages to those consisting of sentences that could possibly be understood, maybe after some thought. Let us follow Loar by calling such languages "effective languages." While this position goes against the tradition to conceive of languages as consisting of a denumerable infinity of sentences, I think one can reasonably take this position. But it is not very popular.

Bennett (1976:§81) argues that it is wrong to restrict languages to effective languages. He prefers the traditional conception for its theoretical "simplicity and smoothness." Hence, he is willing to claim that a population can use such an infinite language. The fact that only a fragment of it is usable is explained by appealing to cognitive limitations of the population's members. Bennett's argument is not very convincing. Plausibly, human linguistic abilities are partly explained by means of a mentally realized grammar (while not necessarily one of the Chomskian sort). If such a grammar is not defined for the very complex sentences, then Bennett's explanation is wrong. But there is another argument against Loar's position, based on an appeal to an intuition many seem to have: If an expression shows certain structural features, then it should be meaningful, even if it is not part of an effective language.

The ones who accept the problem – Lewis included – have proposed different solutions, which can be divided into two groups: *non-mentalistic* and *mentalistic*. Common to them is that they acknowledge that the actual-language relation holds between populations and *effective languages*; what they need then is an addition which determines the meanings of an extension of such a language.

Lewis (1992) and Bennett (1976:§§65–67) suggest a non-mentalistic solution. The proposal is to *extrapolate* a "straight" grammar from the effective language a population uses based on the recurrent features in the history of utterances up to now. Such straight grammars Γ then determine the meanings of actually unused expressions. This is a modification of the elementary theory presented above, one in the step from ALR-S to ALR-E. The relation between a language $\mathcal{L}$ used by a population and the grammar Γ it uses is now as follows. Γ does not (only) generate $\mathcal{L}$ but $\mathcal{L}_\Gamma$ which is a superset of $\mathcal{L}$. $\mathcal{L}$ is an effective language while the language $\mathcal{L}_\Gamma$ generated by Γ is not.

The solution has issues, even when we put worries with respect to a grammar's "straightness" and "bentness" aside. Why should we rely on an external non-mentalistic criterion and buy Lewis' worry according to which otherwise, one would engage in risky speculation about the brain (§6.1.3)? Shouldn't we rather use an internal criterion directly relating to mental properties of the language users? Schiffer (1993, 2006) raised these questions and came up with several counterexamples for Lewis's solution: (i) If members of a population using a language process the utterances not on the basis of a mentally realized grammar but via a huge but finite translation manual into their native language, then they couldn't "go on" and process the sentences in the unused extension of the effective language. In such a case, it would be wrong to claim that these sentences mean something.[12] (ii) What if a population uses a *bent* grammar? In such a case, the Lewisian strategy gives the wrong result. (iii) Lewis' solution requires that there is exactly one straight grammar. But, as also Chomsky (1980:83–85) and Jorgensen (2008) stress, there is no reason for this optimistic assumption. Plausibly, there can be grammars of different frameworks that are straight relative to the framework but don't generate exactly the same language.

[12] O'Leary-Hawthorne (1993) discusses a similar counterexample.

Moreover, the space of possible grammars is huge. For they must draw fine distinctions to explain phenomena like indexicality and vagueness. Every difference in the meaning-assignment yields a different grammar. Hence, a lot of empirical evidence is required to distinguish between different candidate grammars. But the empirical evidence for each language is relatively limited, even on the idealizing assumption that we could obtain it.

Schiffer (1993:244–247) proposed an alternative mentalistic solution. He thinks that the lesson to be learned from his counterexamples is that possessing a mentally realized grammar is relevant in some cases but not necessary. For a French-speaker can use and understand English by using a translation device that translates English utterances into French (and back). But in both cases, Schiffer thinks that we may assume that language users possess a language-processing mechanism which he calls "$\mathcal{L}$-determining translator". Such a translator is something that "determines a mapping of each sentence of a [public abstract] language $\mathcal{L}$ onto a Mentalese sentence that means in Mentalese what the $\mathcal{L}$ sentence means in $\mathcal{L}$" (Schiffer 1993:246). Translators can be realized in different ways, *e.g.* by a mentally realized grammar or a huge translation memory.

Schiffer's (allegedly) "risky speculation about the brain" consists in the following two assumptions: (i) the language of thought hypothesis ("Mentalese") is true and (ii) humans process sentences of a public language by a suitable translator. The strong assumption is (i) since (ii) only requires that there exists a mapping from the public language into Mentalese.

The language of thought hypothesis is one of the central theses of Fodor's computational/representational theory of mind. According to it, (i) mental representations are structured (like the expression "$1 + 2$"); (ii) parts of these structures are "transportable", *i.e.* type-identical parts can occur in different representations (like the "1" in "$1 = 2 - 1$"); and (iii) mental representations have a compositional semantics (*cf.* Fodor 1987:135-137).

The language of thought hypothesis is contested (*cf.* Aydede 2008), as probably all important philosophical theses are. Yet it seems that Lewis overstated his "risky speculation about the brain" point. The hypothesis has a highly general content and is in principle testable: If it turns out that mental representations are not structured and don't have a compositional semantics, then the hypothesis is false. For example, in certain simple neural networks, the hypothesis is false. So what one would have to find out is whether our brain is more like a computer or like a simple neural net (or ...).

Using the notion of a translator, Schiffer's proposal can be stated as follows:[13]

ALR-E′. Expression e means m in $\mathcal{L}_{\mathcal{T}}$ of P iff (i) P uses $\mathcal{L}$, (ii) $\mathcal{L} \subseteq \mathcal{L}_{\mathcal{T}}$, (iii) members of P process $\mathcal{L}$-utterances via an $\mathcal{L}$-determining translator $\mathcal{T}$, (iv) $\mathcal{T}$ generates $\mathcal{L}_{\mathcal{T}}$, and (v) $\mathcal{T}$ pairs e with m.

On this modification of the elementary theory, conventions still play a role: In condition (i) the actual-language relation is used which is defined in terms of conventions. The crucial condition is (iii) which relates the members of P to a translator. The translator is used to generate the extended language $\mathcal{L}_{\mathcal{T}}$ in condition (iv).

If we endorse ALR-E′, then the meaning-determination claim ALR-S is not needed anymore and ALR-E is in general not acceptable. ALR-E can still come out true, namely if the translator is realized by a grammar. In this case, the step from ALR-S to ALR-E is changed: there is no element of choice anymore since the selection of the grammar is achieved by giving it a cognitive role.

Schiffer's solution does not have the problems Lewis's solution has. For example, if the language users process $\mathcal{L}$ via a mentally realized grammar that is "bent," then this grammar (by means of the $\mathcal{L}$-determining translator it realizes) determines the correct meanings.

These three kinds of reactions to the meaning-without-use problem differ in the meanings they assign to unusable expressions. Since Schiffer's solution gives the best result with the fewest problems, I tend to endorse it, even if this commits one to making sense of what it is for an agent to have an $\mathcal{L}$-determining translator.

Consequences The use of translators (and grammars in the typical case) changes the form of the determination claim. Recall that one of the central theses of the conventionalist project is C′ (§1.3):

C′. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable use of e at $\mathcal{C}$

[13] I do not follow Schiffer strictly. He prefers to state his proposal in terms of regularities in speaker-meaning that are mutually known among the members of the respective population. Thereby, linguistic understanding is explained in terms of knowledge about what speakers mean. I prefer my proposal since it's closer to Lewis' and since it does not rely on knowledge.

determines that e means m at $\mathcal{C}$.

I introduced the thesis in a noncommittal way and mentioned the possibility that its scope would have to be restricted. Indeed, the scope has to be restricted: (i) There is meaning without use. (ii) We need to assign meanings to sub-sentential expressions. For these reasons, a change of the explanatory architecture is required. Only the meanings of sentences of effective languages (the usable fragments) can be said to be determined by conventions. The meanings of unused expressions and sub-sentential expressions are determined by a translator.

This change in Lewis' account raises two questions I'd like to consider: On the face of it, why do we invoke conventions at all and not only, for example, Schiffer's $\mathcal{L}$-determining translator? What remains of the conventionalist thesis?

Schiffer's solution should make us aware of alternative "unconventionalist" explanations of why (in virtue of what) expressions mean what they do. It seems that what explains an expression's meaning is now something individual – an $\mathcal{L}$-determining translator – and not something social anymore – a convention or social norm. This observation is only in part correct. Indeed, some of the explanatory work is now done by $\mathcal{L}$-determining translators. Could the translators do all the work? Let us consider an account that tries to explain linguistic meaning solely in virtue of *someone* possessing a translator. My claim is not that this is a position someone would want to endorse but it's an instructive proposal. For such an account would fail badly and if we improved on it, we would have to add conventions and social norms.

As to its failure: (i) Endorsing such an account is to accept Humpty Dumpty's claim that he can make a word mean what he wants. Since we reject his claim, we should generally judge so (with some exceptions relating to social norms, see §9.3 and §9.4). (ii) We cannot explain why communication is usually easy and successful solely on the basis of the assumption that agents are individually able to use and understand an idiolect (say by means of an $\mathcal{L}$-determining translator). This was one of the lessons learned from the discussion of Davidson's argument (§3.3). Conventionalist accounts provide a very plausible explanation of this feat: It's because there is a conventionally shared language among a population of which the communicators are part. (iii) The possession of an $\mathcal{L}$-determining translator secures neither the arbitrariness of meaning, its stability, nor that it's beneficial for each communicator to coordinate (since no other communicators are considered in the first place). (iv) It seems to be impossible to explain semantic normativity on such an account.

Notice that these are all features that are part and parcel of a conventionalist account (with an account of social norms). The second part of my claim should now be plausible: if we tried to improve the account under consideration, then we would have to add things which would amount to adding conventions and social norms.

Thus, the question is not: "Should we abandon the conventionalist project (since we incorporate grammars/translators now)?" but rather: "What's the relation between translators and conventions?" I think the following is a plausible beginning of a more complete answer: As before, the explanation starts with something social: a stable linguistic use like a convention (or a social norm). From the existence of such a stable linguistic use it follows that its members have certain dispositions to use and understand certain expressions. This was the conventionalist part of the explanation. The crucial claim to relate it to the grammar part is the following *stipulation*. These dispositions are brought about, at least in part, by an $\mathcal{L}$-determining translator. To make this stipulation explicit, let us say that *linguistic conventions* are conventions whose members' behavioral dispositions are brought about, at least in part, by translators, and likewise for "linguistic social norms" and "linguistic normative conventions" that are yet to be introduced. Hence, translators are constitutive for linguistic conventions. That is to say: while $\mathcal{L}$-determining translators have an explanatory role in the overall explanation, the explanation is not just in terms of them.

To conclude, the conventionality thesis C′ has to be restated since the existence of a convention does not imply that there is a grammar which is needed to address the meaning-without-use problem and to assign meanings to sub-sentential expressions. If we introduce the term "stable linguistic use" for linguistic conventions and linguistic social norms (and combinations of them), then we can formulate the thesis with a more cognitive underpinning as follows:

C″. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable linguistic use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

In the case of meaning in virtue of conventions, we can be more specific: According to a Lewisian Actual Language Relation theory, an expression's meaning is determined by the $\mathcal{L}$-determining translator that is used to process the effective language used (ALR-E$'$). Conventions determine the meanings of the sentences of the effective language. Translators determine the meanings of the extended language and of all sub-sentential expressions.

6.1.6 Language varieties

The orderliness of Lewis' theory is suspicious. Can it really be that simple? Language varieties have been considered to be problematic for Lewis' account. But is that so? Before I discuss Jorgensen's objection in the evaluation-section (§6.2.1), I elaborate on Lewis' proposal.

A first observation is that the folk ways of individuating languages are quite different from Lewis' way. His abstract languages have very precise identity conditions: one little change d, say a more word or less, yields another formal language – $\mathcal{L} \pm d \neq \mathcal{L}$. Natural languages are vague objects: a little change does not yield another language – $\mathcal{L} \pm d \approx \mathcal{L}$.

This modeling mismatch is not problematic on its own. It suggests, however, that we shouldn't translate ordinary language sentences of the kind "the Germans use German" too directly into the theoretical vocabulary as "population P_G uses the abstract language $\mathcal{L}_G$." For there are many varieties of German, some of them blending neatly into Dutch and others into Austrian. We can describe these varieties as a family of abstract languages that have shared parts and overlapping parts in which they disagree (say one language assigns to an expression one meaning and another assigns a different meaning to it). This raises the question which of the many abstract languages $\mathcal{L}_G$ should designate. Luckily, since a population can use more than one abstract language, we don't have to choose exactly one (*cf.* Lewis 1975:397 ff.). It follows from Lewis' theory that, *ceteris paribus*, if a population uses an abstract language $\mathcal{L}$, then it also uses all abstract languages that are a subset of $\mathcal{L}$.[14] Hence, the cases of interest are ones where the abstract languages disagree on the meanings they assign to the shared expressions. Lewis considered two particular kinds of languages in case of a linguistically inhomogeneous population, namely its *shared language* and its *total language*, the former being the *intersection* of the abstract languages

[14] At least on the assumption that members of a population don't lose interest in a language as it gets smaller.

that are used by any substantial subpopulation and the latter being its *union*. I think the former notion is more interesting since it allows us to explain why communication is easy in a population.

It might not be obvious which kind of regularities have to exist in such population for it to use a particular shared language. For in the case we're considering now, the regularities of truthfulness and trust might be in different languages among many members and there are occasional non-conforming uses. So, let me illustrate Lewis' idea by discussing the simplest scenario: there are two distinct expressions e_1 and e_2 and two distinct meanings m_1 and m_2, defining the eight possible languages in figure 6.1(a). Figure 6.1(b) on the right shows the intersections of the languages of the respective row and the respective column.

$$\begin{aligned}
\mathcal{L}_1 &= \{\langle e_1, m_1\rangle\}\\
\mathcal{L}_2 &= \{\langle e_1, m_2\rangle\}\\
\mathcal{L}_3 &= \{\langle e_2, m_1\rangle\}\\
\mathcal{L}_4 &= \{\langle e_2, m_2\rangle\}\\
\mathcal{L}_5 &= \{\langle e_1, m_1\rangle, \langle e_2, m_1\rangle\}\\
\mathcal{L}_6 &= \{\langle e_1, m_1\rangle, \langle e_2, m_2\rangle\}\\
\mathcal{L}_7 &= \{\langle e_1, m_2\rangle, \langle e_2, m_1\rangle\}\\
\mathcal{L}_8 &= \{\langle e_1, m_2\rangle, \langle e_2, m_2\rangle\}
\end{aligned}$$

(a) Languages

$\bigcap$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_7$	$\mathcal{L}_8$
$\mathcal{L}_1$	$\mathcal{L}_1$	$\emptyset$	$\emptyset$	$\emptyset$	$\mathcal{L}_1$	$\mathcal{L}_1$	$\emptyset$	$\emptyset$
$\mathcal{L}_2$		$\mathcal{L}_2$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\mathcal{L}_2$	$\mathcal{L}_2$
$\mathcal{L}_3$			$\mathcal{L}_3$	$\emptyset$	$\mathcal{L}_3$	$\emptyset$	$\mathcal{L}_3$	$\emptyset$
$\mathcal{L}_4$				$\mathcal{L}_4$	$\emptyset$	$\mathcal{L}_4$	$\emptyset$	$\mathcal{L}_4$
$\mathcal{L}_5$					$\mathcal{L}_5$	$\mathcal{L}_1$	$\mathcal{L}_3$	$\emptyset$
$\mathcal{L}_6$						$\mathcal{L}_6$	$\emptyset$	$\mathcal{L}_4$
$\mathcal{L}_7$							$\mathcal{L}_7$	$\mathcal{L}_2$
$\mathcal{L}_8$								$\mathcal{L}_8$

(b) Shared languages

Figure 6.1: A simple scenario of language varieties

Let us consider a population consisting of three agents, being in the position to be truthful and trusting in one of the eight languages (or in short (individually) "using" one the languages).[15] Thereby we can simulate scenarios in bigger populations where the agents are subpopulations and the languages are much bigger. For each agent, there are eight languages she can use. (If she is using more than one compatible language, say $\mathcal{L}_1$ and $\mathcal{L}_4$, then we count this as using one language, namely its union $\mathcal{L}_6$. Languages like $\mathcal{L}_1$ and $\mathcal{L}_2$ are incompatible in this sense since the same expression has two different meanings and we assume that one cannot be truthful and trust-

[15] An agent's being "truthful and trusting in $\mathcal{L}$" has to be defined in the obvious way. I omit here complications resulting from agents understanding languages they can't speak and the like.

ing in both of them. Consequently, there are already 256 ($= 8^3$) possible configurations resulting from the agents using one of the languages. We can characterize the configurations by using a three-dimensional coordinate system where each point is a triple consisting of the three languages used by the agents, respectively. Obviously, in configurations of the form $\langle \mathcal{L}_i, \mathcal{L}_i, \mathcal{L}_i \rangle$, the population is using $\mathcal{L}_i$. This is the linguistically homogeneous case.

Now let us turn to cases of linguistically inhomogeneous populations. Lewis seems to suggest the following procedure to determine the shared language of a population: First determine the languages used by each substantial subpopulation (and make sure that these languages are not too different yielding an empty intersection). Then take the biggest language of each such subpopulation. The union of these languages is the shared language of the population.

Thus, we might think that based on Lewis' proposal the population simply uses $\mathcal{L}_\cap$ in configurations of the form $\langle \mathcal{L}_i, \mathcal{L}_j, \mathcal{L}_k \rangle$ where $\mathcal{L}_\cap = \mathcal{L}_i \cap \mathcal{L}_j \cap \mathcal{L}_k$. But since Lewis allows for some exceptions in the conventional regularities, the story is more complicated. Let us just stipulate that if two of three agents use a language, then the whole population uses it. Thus, a population is also using $\mathcal{L}_i$ in configurations of either the form $\langle \mathcal{L}_i, \mathcal{L}_i, \mathcal{L}_k \rangle$ or $\langle \mathcal{L}_i, \mathcal{L}_k, \mathcal{L}_i \rangle$ or $\langle \mathcal{L}_k, \mathcal{L}_i, \mathcal{L}_i \rangle$.

Less obviously, there is also a case where one of the used languages overlaps another used language. In such cases, if there is a non-empty intersection-language of two used languages, then it is also used by the population. That is, even if the intersection-language $\mathcal{L}_\cap$ of all three agents is empty and no two languages are identical, there are configurations where the population uses a non-empty language. To illustrate this, let us enrich the vocabulary by two more expressions and meanings. Then there is a configuration where agent 1 uses $\{\langle e_1, m_1 \rangle, \langle e_2, m_2 \rangle\}$, agent 2 uses $\{\langle e_2, m_2 \rangle, \langle e_3, m_3 \rangle\}$, and agent 3 uses $\{\langle e_3, m_3 \rangle, \langle e_4, m_4 \rangle\}$. In this configuration, the intersection-language $\mathcal{L}_\cap$ of all agents is empty and no two agents are using the same language. But there are the intersection languages $\{\langle e_2, m_2 \rangle\}$ and $\{\langle e_3, m_3 \rangle\}$ of at least two agents. By allowing exceptions to the regularity, these languages are also used by the population.

So, Lewis' theory entails that in certain cases of overlapping abstract languages that disagree on the meanings they assign to the shared expressions, the overlapping part cannot be a (sub-)language used by the whole population. Intuitively, this seems plausible for the members of the whole population wouldn't communicate successfully by using this sub-language. This does not rule out that the conflicting parts of the two abstract-languages can be used by the respective sub-populations of the population: As long as there is a regular not too conflicting use of them, the parts are actually used.

To conclude, already in the simplest scenario, the application of Lewis' theory is complicated. It gets more difficult as we add recursive rules, more agents, and changes over time. This is a disadvantage of Lewis' theory since the standard case among humans is one of linguistically inhomogeneous populations.

6.2 Evaluation

Even with the complications stemming from the meaning-without-use problem, Lewis' Actual Language Relation theory has obvious appeal: it is simple and has relatively well understood foundations. On the basis of the presentation and discussion of Lewis' theory, we can now evaluate it. I skip a detailed discussion of the adequacy conditions (§1.4) since the outcome is almost the same as for his Signaling Games theory (§5.4). This is so because the same account of conventions is used and because the conventional regularities are similar.[16] With regard to the adequacy conditions there is one improvement: the new theory can explain linguistic structure and thereby linguistic meaning (DesM6). For the same reason, the new theory inherits also the problems of Lewis' Signaling Games theory, namely it's rationalistic conception of language use and its users, on the one hand, and the lack of an account of semantic normativity, on the other hand. Thus, meaning is not normative on Lewis' theory while it should be possible that it is (chapter 2). Lewis (vaguely) suggested to add social norms on top of conventions: If there is a social norm to conform to a convention, then there are *oughts* (*c.f* Lewis 2002:97–100).[17] Since he does not provide an account of social norms, his theory is incomplete with respect to semantic normativity and with respect to meaning in virtue of social norms. A further consequence of this is that he is stuck with implausible mood explications – a plausible

[16] But not identical: the direction of the conditional in the speaker's portion of the regularity is reversed from "belief → utterance" to "utterance → belief."

[17] But this would go against the principle promoted in chapter 2 that if meaning is normative, then this is because of the way it is determined.

explication of what it is *to inform someone* is in terms of social norms (§2.5).

In this section, I discuss some further criticism. I start by discussing four objections against Lewis' theory: an objection by Jorgensen drawing on language varieties (§6.2.1), an objection based on Davidsonian language use (§6.2.2), an objection by Max Kölbel advocating an alternative definition of the actual-language relation (§6.2.3), and an objection based on the Gricean heritage of Lewis' theory (§6.2.4). In §6.2.5, an interim summary is given and Gricean variations are explored, in particular alternative Actual Language Relation theories are considered.

6.2.1 Jorgensen's objection from language varieties

Andrew Jorgensen (2008:82 ff.) argues on the basis of language varieties that in cases of linguistically inhomogeneous populations there is no language the population uses:

> Consider the case of English. The body of sentences and rules that constitutes the "lowest common denominator" variety of English understood by all speakers of English must be sheared of all the peculiarities of Hiberno-English, Ulster Scots, Lallans, Doric, Strine, Newziln, Yorkshire English, Mancunian, black English vernacular, and so on. [. . .] [T]here is a plurality of natural ways of extrapolating the used part of this structure to the unused cases. It is thus the case that existing linguistic diversity shows that condition 3 [requiring that everyone has a good and decisive reason to conform to one of these fragments] fails for these proper language fragment-type cases. This conclusion can be extended. If condition 3 fails for the proper fragments of a language, then it fails for languages themselves at a time, since the entire corpus of a language up to a certain moment in time is a proper fragment of the entire corpus of a language up to some succeeding moment.
>
> (Jorgensen 2008:83)

If I understand his argument correctly, then it can be stated as follows:

P1. In cases of a linguistically inhomogeneous population, there are many language varieties used by the subpopulations. These language varieties are (finite) effective languages. For each language variety, there are many admissible extrapolated languages.

P2. A population uses an abstract language $\mathcal{L}$ only if each member's belief that the others conform to the regularity of truthfulness and trust in $\mathcal{L}$ gives her a *good and decisive* reason to conform to the regularity herself.

P3. But in such a case, there is no such *good and decisive* reason for any member of the population to be truthful and trusting in one variety since she could equally well be trusting in another variety.

C. The population uses no language.

It's clear that Jorgensen is on to something. For we have seen in §6.1.6 that Lewis' theory becomes opaque when applied to linguistically inhomogeneous populations. Nevertheless, I think Jorgensen's argument is not sound.

Observe that Jorgensen's point is not that there is no shared language. He seems to grant that there might be a non-empty intersection-language ("lowest common denominator"). This language would be the candidate Lewis' theory selects as (one of) the language(s) used by the population. Jorgensen's objection seems to result from the combination of (i) there being many extrapolated languages and (ii) Lewis' necessary condition 3 of a convention. Condition 3 requires that the belief in the others' conformity gives each member a good and decisive reason to conform to it themselves (see above in §6.1). Jorgensen's point then seems to be: Each member of the population faces a non-trivial decision problem. There are many candidate languages one could be truthful and trusting in. But contra Lewis, there is no good and decisive reason to choose one since many of them are equally good.

Still, I think that the argument is not very compelling since P3 seems to me based on a misunderstanding. It seems to me that Jorgensen implicitly assumes that there has to be exactly one language which has to be extracted from all the language varieties in the currency of a linguistically inhomogeneous population and this language is the best choice no matter what. This is wrong. First, by endorsing Schiffer's solution for the meaning-without-use problem, each variety should have only one interesting extension and thus the number of candidates to choose from is not as big as Jorgensen thinks. Second, in such a case, Lewis' proposal is not as Jorgensen thinks: We start by considering the biggest effective languages used by substantial subpopulations. By taking the intersection of these languages, we construct the shared language of the population. There is only one such language. Third, each individual of the population does not face only *one* decision problem with regard to the language to use but in fact many. For an individual can be a member of different subpopulations using different languages. The language she should use depends on the subpopulation she's a member of.

Crucially, in each subpopulation, there is a language for whose use she has a good and decisive reason. Finally, having no further knowledge than knowing to be a member of a linguistically inhomogeneous population, using the shared language seems to me the rational thing to do. For this reason, Lewis' proposal is recommendable.

To conclude, Jorgensen's argument does not have the force he thought it to have. Nevertheless, linguistically inhomogeneous populations seems to be the standard rather than the exception. The discussion shows that the application of Lewis' theory is complicated in such scenarios. This is so because Lewis tries to relate two abstract entities by means of abstract conventions: populations and abstract languages. Both are abstract entities which are intended to model vague worldly counterparts: linguistic communities and living languages. Seen this way, the complications should not surprise us. For at the heart of the complications is a deviation from our pre-theoretic conception which I take to be that there is a convention for each word. That there are these complications is a point against Lewis' theory. An account that could deal with such scenarios in a better way is obviously to be preferred.

6.2.2 Sperber and Wilson's objection from Davidsonian uses

Lewis' regularities of truthfulness and trust suggest that communicators don't lie and use language literally all the time. If so, isn't Lewis' theory at odds with reality? While Lewis does not discuss malapropisms, he discusses inveterate liars, hyperbole, metaphors, joking language uses, and Gricean communication (Lewis 1975:395–397). He offers different explanations of the phenomena. Let me first present Lewis' position before I evaluate it.

The case of inveterate liars is quickly told: Lewis denies that a population of inveterate liars can use a language at all. As another case of seeming untruthfulness, he considers the use of hyperboles and metaphors. Here the claim is that such language use is one conforming to a regularity of truthfulness and trust, in spite of all appearances to the contrary. As Sperber and Wilson (2002:587–589) explain, Lewis seems to suggest that a sentence like (1) is ambiguous, having besides its literal meaning the same meaning as (2).[18]

(1) You are a piglet.

[18] The example is from (Sperber and Wilson 1986:154).

(2) You are a dirty child.

For joking and more broadly insincere language use, Lewis does not offer just one treatment but three. The first one is as above: The seeming untruthfulness in the language is actually truthfulness in another closely related language. In the language the people use they are truthful but they are untruthful in a closely related more systematic literal-language. In the language they use, the insincere uses of the sentences are indicated by signs of insincerity (example: child speaks with crossed fingers).[19]

The second treatment is to count the insincere uses as *exceptions* to being truthful and trusting. The third treatment is by a modification of Lewis' theory: the relevant language use that determines the language(s) used by a population is restricted to the use in *serious communication situations*.[20] Plausibly, malapropisms could be explained along similar lines.

Finally, Gricean communication (communication obeying the Gricean maxims) is not in conflict with his theory. Social norms of being helpful and relevant can exist together with conventions of truthfulness and trust in a language.

A look in the literature suggests that Lewis' claims are not generally accepted. Davis (2003:289 ff.) finds cases of inveterate liars using a language conceivable since these liars use the words in their normal meanings. Sperber and Wilson (1986:154) point out that long stretches of literal language use are not the rule but exceptional and argue that (i) non-literal language use and understanding is not based on literal use and understanding and that (ii) – contrary to Lewis – non-literal meanings are a different kind of meaning than literal meanings since the non-literal meanings of figurative utterances of expressions are highly context dependent.

I myself don't find it plausible to say that inveterate liars can use a language in the ordinary sense. While I see Davis' point, I think that it only applies to a short transition period in the liar community. If conventions determine the words' meanings, then after a while, their words will have

[19] I think that this proposal is quite strong. For recent research suggests that heuristics combined with simple machine learning is sufficient to reliably recognize sarcasm in closed domains: Tsur et al. (2010) developed a sarcasm recognizer that recognizes 77% of sarcastic sentences in the domain of Amazon customer reviews.

[20] Lewis defines such a situation as one in which it is common knowledge that the hearer wants to know whether some sentence s is true and in which the speaker knows that; moreover, the speaker is assumed not to have a strong preference for uttering s or not.

lost their meanings since the conventions will cease to exist (due to the lack of a common interest to communicate).

Sperber and Wilson's criticism is more difficult to assess since it bears on contested topics in the current semantics/pragmatics debate. For example, in case of example (1), it seems to me quite plausible to say that one of its literal meanings is (nearly) synonymous to (2). In other cases, it seems to me more plausible to explore (while not necessarily to endorse) the claims of Sperber and Wilson. The problem seems to be how the literal meanings are put to use in communication. Lewis suggests a very direct connection. The radically pragmatic theory developed by Sperber and Wilson (1995) assumes a very indirect connection resulting in a very weak role for literal meanings. Even together with indexical information, such meanings for sentences are not truth-conditional. A more moderate proposal is offered by Recanati (2004) who introduces an optional semantic "modulation" process which maps semantic values of expressions to other semantic values of the same type (not freely, of course, but the constraints are contested).

While Sperber and Wilson's proposal seems to be incompatible with Lewis' theory, Recanati's modulation seems to be compatible. For this reason, I think that one should work out the details of such an account in the semantics/pragmatics debate.[21] I'd like to suggest that the regularities of truthfulness and trust are redefined as follows: The truth-conditional-element of an utterance of an expression is, depending on the mood, the *modulated meaning* of the truth conditional-element of the uttered expression.

So, while these considerations are sketchy and inconclusive, they seem to point to a weakness of Lewis' account. But, contrary to Sperber and Wilson, I think that one is not forced to endorse their account. It seems to me that one could keep most of Lewis' theory by incorporating Recanati's modulation.

6.2.3 Alternative regularities: Kölbel's lust and lies

Kölbel considers in his article *Lewis, language, lust and lies* (Kölbel 1998) alternatives to Lewis' regularities of truthfulness and trust, arguing for a change in Lewis' theory using a disjunctive regularity. Since no one has

[21] Recently, Pagin and Pelletier (2007:§4) have developed a compositional semantics for Recanati's proposal. It highlights the need to impose constraints on modulation.

challenged his argument, I'll do it now. Among the regularities Kölbel considers are:

1. a regularity of *reliability and reliance* in a language $\mathcal{L}$ which is to reliably making only true-in-$\mathcal{L}$ utterances and to rely on the truth-in-$\mathcal{L}$ of utterances (p. 304)
2. a regularity of *sincerity and trust* which is to make an utterance only if one believes it to be true-in-$\mathcal{L}$ and to respond to utterances by coming to believe that their speaker believes them to be true-in-$\mathcal{L}$ (p. 305)
3. a regularity of *cageyness* which is to make an utterance only if one wants to give the impression that one believes the utterance to be true-in–$\mathcal{L}$ and to respond to utterances by coming to believe that their speaker wants to give the impression that he believes them to be true-in-$\mathcal{L}$ (*cf.* p. 309)

Kölbel argues that a regularity of sincerity and trust cannot explain the flow of information. For if Romeo utters "ti amo," addressing Juliet, then Juliet comes to believe that Romeo believes his utterance to be true in Italian but not *that Romeo loves Juliet*, as the regularity of reliability and reliance has it. The former would need additional assumptions to explain the flow of information. Kölbel takes this to be a reason for the latter regularity.

But if Lewis' theory is understood in this way, then it has problems with scenarios where the speaker is unreliable. Kölbel (1998:304–308) argues that if such scenarios occur frequently, then no regularity of reliability and reliance can be a convention. One reason is that there are too many exceptions and Kölbel believes that Lewis could not allow for more deviations. The second reason is that common knowledge is precluded in such a scenario.

Kölbel (1998:308–312) iterates this point: In case of overt lies, the communicators plausibly "see through" them. In this case, it's more plausible that there is a regularity of cageyness since the communicators would not be sincere and trusting.[22] The disadvantage of postulating such a conventional regularity is – as in the case of sincerity and trust – that the flow of information is also not explained. It seems that the additional assumptions need to be even stronger than in the case of sincerity and trust.

Conceivably, a speaker could be even more cagey by just wanting to give the impression that . . . she wants to give such an impression, *etc.* Since we

[22] Also Williams (2002:70) and Laurence (1996:278 ff.) argue against a regularity of truthfulness and trust on the basis that we utter sentences we believe not to be true.

want one kind of regularity to define the actual-language relation, Kölbel thinks that we should define the regularity as a disjunction, roughly as follows: A population uses a language iff among them, there is regularity in *reliability and reliance* or *sincerity and trust* or *cageyness* or ... (for more cagey regularities).

Let me turn now to the evaluation of Kölbel's argument.[23] Kölbel is wrong in thinking that Lewis couldn't weaken the degree of conformity to a regularity. I've argued that Lewis' theory of conventions could be modified by taking an *expected-utility* approach allowing for conventions with widespread deviations (§4.3.3). Parties to a convention in this sense only have to conform often enough, where "often enough" is defined in terms of expected utilities. As to common knowledge, I think we've heard by now enough reasons not to insist on it. On the basis of these rejoinders, we can block Kölbel's moves and are not forced to accept his proposal in terms of a disjunctive regularity. Moreover, I don't share all of Kölbel's intuitions. Why isn't there a flow of information in case of a convention of sincerity and trust? After all, the hearer learns something about the speaker and the additional assumption that normally the speaker speaks truly is also plausible. To conclude, I agree with Kölbel that Lewis' theory needs a modification. We disagree about the place. Kölbel wants to redefine the actual-language relation; I prefer to redefine the notion of a convention since the root of the problem seems to me more general than language use: conventions should allow for a substantial share of deviating behavior.

6.2.4 Being Gricean is a handicap

Actual Language Relation accounts are typically Gricean in the sense that if a speaker utters something conforming to a linguistic convention, then the speaker means something – in Paul Grice's sense of "speaker-meaning". I turn now to the question whether this is a desirable feature or not. The target is not speaker-meaning as Grice originally analyzed it in his article *Meaning* (Grice 1957) but a version based on his later article *Utterer's meaning and intention* (Grice 1969) which I formulate as follows (for simplicity restricted to indicative utterances):[24]

[23] I think it's unfortunate that Kölbel has chosen the case of love since, arguably, we all don't know very well what "love" means.

[24] This formulation is based on (Neale 1992:550) but without dropping clause 3 which I think is part of the core of Grice's proposal.

For all speakers S, utterances u, and propositions p: By uttering u, S meant that p iff for some hearer H,

1. S uttered u intending H actively to entertain the thought that p (or the thought that S believes that p);
2. S uttered u intending (R) H to recognize that S intends H actively to entertain the thought that p;
3. S uttered u intending R to function, in part, as a reason for H to recognize that S uttered u intending H actively to entertain the thought that p (or the thought that S believes that p) (=1.);
4. S does not intend H to be deceived about S's intentions 1. and 2.

The original analysis in *Meaning* had a couple of problems I don't want to rehearse here in detail.[25] For example, the intended reaction of the hearer in clause 1 was initially *to believe that p*. But in exam-like situations the examiner already has the relevant belief and the examinee expects that the examiner has this belief. Another problem were argumentative dialogs where the speaker doesn't want the hearer to come to accept the conclusion because she intends so. Also, deceptive intentions (now ruled out by clause 4) have been a matter of some controversy. The important point for our discussion is that Griceans would endorse something like the new analysis presented above and that this version does not have the well-known problems of the original analysis. At least, I will grant this to the Griceans in order to focus on two issues I find most important: The sentential primacy thesis presupposed by Gricean analyses and the Intellectualism they entail.

Sentential primacy That sentences are in some sense prior to sub-sentential expressions is a widespread assumption, endorsed in particular by broadly Gricean analyses of meaning (and not only Paul Grice's).

I'll argue that in a certain sense such an assumption is problematic. To do so, (i) I clarify the relevant sense of priority (or primacy), (ii) I relate the sentential primacy thesis SP to broadly Gricean analyses, and (iii) I provide reasons against SP.

The sentential primacy thesis that is relevant for the conventionalist project is the one which concerns the determination of sub-sentential meanings (since the project is concerned with their determination):

[25] Grice (1969) discusses many cases. For extensive discussion, consult for example (Kemmerling 1979, 1986; Neale 1992; Davis 2003, 2005; Schulte 2008).

SP. Sub-sentential expressions have the meanings they do *because* sentences in which they occur mean what they do.

SP is a thesis expressing how meanings of sub-sentential expressions are *determined* (in the sense of "determination" introduced in §1.1.2 which amounts to something stronger than global supervenience – *e.g.* to conceptual entailment as I prefer). If one endorses SP, then one claims that sub-sentential expressions mean what they do *solely* in virtue of what the sentences mean in which they occur.[26]

SP is *not* an epistemological thesis[27] as some understand it.[28] For this reason, the truth of the epistemological versions of SP is irrelevant for my discussion.[29]

Moreover, we should understand the terms "word", "sub-sentential expression", and "sentence" in their syntactic sense and semantic sense. The syntactic sense is fixed by paradigm cases of what we count as *words*, *sub-sentential expressions*, and *sentences*. Linguistic tests help us to determine the syntactic categories of expressions in a systematic way, in particular the movement-test according to which two expressions are of the same syntactic category if they can occur in the same environments. For example "a farmer" is not of the same category as "it rains" since in the environment "He said that ___ and after the rain comes sun", "a farmer" cannot be substituted for "___" while "it rains" can. Such tests are not infallible but still useful since they are reliable to some degree.

Also the semantic senses of "word", "sub-sentential expression", and "sentence" are fixed by paradigm cases. Sentential expressions have a truth value in context while the semantic values of sub-sentential expressions in context are not propositional.

The syntactic and semantic senses of "word", "sub-sentential expression", and "sentence" contrast with a pragmatic sense of "sentences" according to which they are the primary unit used to perform a speech act.[30] For the discussion, this pragmatic sense is not at stake.

Broadly Gricean analyses of meaning presuppose SP but we have good

[26] See (Szabó 2008:402–403) for a helpful clarification of the thesis.

[27] *Cf.* (Davis 2003:178).

[28] See (Stainton 2006:§10).

[29] One could even grant the truth of a thesis along the lines of "The meanings of sentences are learned before the meanings of words are learned."

[30] See (Stainton 2006), especially §10.1.

reasons not to endorse it.[31] Let me elaborate.

According to Gricean analyses of meaning, an expression *e* means that *p* just in case people use *e* to mean that *p* (Davis 2003:167). On the basis of this definition (be it an analysis or explication), sub-sentential expressions cannot be assigned a meaning. Consider "red": People don't use it to mean *that red* (p. 174).

One could just deny that sub-sentential expressions have a meaning. But this move has consequences that are not worth it:

> It is implausible enough to claim that "blue" and "green" are like "elub" and "neerg" in being meaningless, but it is particularly hard to accept that "The sky is blue" is meaningful but not "the blue sky." Worse yet, the nihilist conclusion would undermine the fundamental principle of the compositionality of meaning. Because "gleft" is meaningless, any phrase, clause, or sentence in which "gleft" is used rather than mentioned is meaningless. So either the nihilist conclusion about [the meaning of sub-sentential expressions] would spread to sentences, or else the Gricean would face the difficult task of saying how "All boys are gleft" differs from "All boys are male." (Davis 2003:175)

So, from a Gricean perspective, there seems to be no other choice than (i) to restrict the analysis of meaning to sentences and (ii) to endorse SP. This is what Griceans do, *e.g.* (Grice 1989d:128–137), (Schiffer 1972:6,166), and (Bennett 1976:16–22, 212–221,272–276,280–284).[32]

Stephen Schiffer, for example, writes in his introduction to *Meaning* in a section called "An order of priorities" about the definitional order of the analysis of sentential and sub-sentential meanings:

> What can be said about the order of priorities obtaining between our [...] analysanda? [...] A prima facie reason for thinking that the concept of a whole-utterance type cannot be logically prior to the concept of a part-utterance type is that the meaning of a sentence is partly a function of the meaning of its words. On the other hand, there are stronger reasons for thinking that the notion of a part-utterance-type [a sentential expression] is logically prior to the notion of a part-utterance type [a sub-sentential expression]. (Schiffer 1972:6)

In this passage, Schiffer quickly considers an alternative to SP and then rejects it in favor of SP by assuming that the notion of a sentential mean-

[31] I base my discussion on (Davis 2003:§8.4) and (Stainton 2006:§10).

[32] See (Davis 2003:175 fn. 16) for many more references.

ing comes earlier in the sequence of definitions provided by a foundational theory of meaning.

Thereby, sub-sentential meanings are defined in terms of sentential meanings. This is a (strong) version of SP since the determination claim expressed by SP follows from it. This way of defining the meanings of sub-sentential expressions is characteristic for broadly Gricean accounts. But there are at least five reasons against SP:[33]

First, the restriction is unnatural. It is natural to say things like "By 'bleu' I mean *blue*" without giving an indirect analysis for these phrases. SP creates a disunity of the analysis by treating sub-sentential expressions differently than sentential expressions – or as Davis puts it: SP "postulates diversity where we should expect uniformity" (p. 181). I think Davis' point is rather an expression of an intuition (which I find plausible) than a decisive objection; for the objection is merely conditional: If you share my intuition, then ...

Second, by accepting SP one endorses a commitment to explain how the meaning of sub-sentential expressions are to be derived from sentential expressions. Otherwise, the meanings of sub-sentential expressions are unexplained and this would be a serious defect of a conventionalist account of meaning (and of foundational theories of meaning in general). But so far, Griceans are vague about this step and the known approaches in the Gricean literature have problems. Hence, unless a plausible derivation procedure is provided, one should not endorse the sentential primacy thesis (Davis 2003:182–187). Moreover, such a procedure faces the following fundamental indeterminacy problem.

Third, Lewis (1975) pointed out – informed by the indeterminacy of meaning arguments of Quine (1960) – that the meanings of sub-sentential expressions are highly indeterminate. For any kind of grammatical description for the sub-sentential expressions is admissible as long as it yields the same distribution of sentential meanings. By complicating the grammar, some words need not be described as words. Also the meanings of sub-sentential expressions can vary wildly as long as the sentential meanings resulting from their composition remains unchanged. This is bad since an account that can assign meanings to sub-sentential expressions more directly and in a less indeterminate way is preferable.

[33] Davis (2003:176–180) also convincingly argues against standard arguments in favor of SP; Szabó (2008:403) shares this judgment.

Davis (2003:187) illustrates the point by discussing a “toy” language which has more words than sentences, namely the three expressions “a”, “b”, and “L”, and the two sentences “aLb”, and “bLa”. In this case, the meanings of the sub-sentential expressions cannot be uniquely determined. For there are three variables (the expressions) whose values have to be determined by two equations (the sentences).[34]

The force of the indeterminacy objection comes from the fact that various indeterminacy claims can be proved. Notably, Hilary Putnam (1983) applied the Löwenheim-Skolem theorem to show that any language that has an extensional semantics, has many different extensional semantics that result from reshuffling the semantic values of sub-sentential parts.[35] In (Putnam 1981:§2), Putnam generalized the “if there is one, there are many” result to intensional semantics.

There are strategies to deal with these indeterminacies (see Janssen 1997): One has to impose constraints on (i) the structure of the language, (ii) the properties of the semantics (*e.g.* what counts as an admissible semantic value for a property-term of the language), and (iii) the admissible semantic mappings. But this is not coherent with the sentential primacy thesis. In effect, by following the strategy, one makes assumptions about what sub-sentential expressions mean.

Fourth, words and phrases can be used and understood in isolation, *e.g.* (i) to make assertions, as in utterances of “John’s father” to make the assertion that the man near the door is the father of John Adams (Stainton 1993, 1994); (ii) on shopping lists, CDs cover, dictionary entries, menus, . . . (Stainton 2004); or (iii) as discourse initial fragments, as in “a beer” to order a beer from the barkeeper (Staudacher 2007).

The use and understanding of linguistic expressions determines their meanings (§1.3). For this reason, the above examples are relevant. They provide a good reason to suppose that also such uses and understandings determine their meanings.

[34] On pp. 187–189, Davis (2003) generalizes his indeterminacy-claim to richer languages. But as Szabó (2008:403) points out, it seems to be an open question for which class of richer languages the claim holds.

[35] Putnam intended his so-called “model-theoretic argument” as a refutation of scientific realism. This claim is by now widely contested. Moreover, Putnams’ proofs have technical problems. The philosophical assumptions of his argumentation have been found implausible. Nevertheless, semantic indeterminacy is usually acknowledged. See (Bays 2001) and (Bays 2009:§3.3.) for a recent discussion.

Observe that the sub-sentential expressions in these examples cannot be analyzed as ellipses in the sense of syntactic or semantic ellipsis theories. Syntactic ellipsis theories assume that ellipses are sentences in the syntactic sense that can be reconstructed from the linguistic discourse context (*e.g.* by using sentential parts from the sentences that have been uttered before the utterance of the ellipsis). Semantic ellipsis theories assume that ellipses are sentences in the semantic sense whose sentential semantic value can be reconstructed from the semantic values of the expressions uttered before the ellipsis in the linguistic discourse context.

But the sub-sentential expressions above can occur as discourse initial fragments. Hence, there is no linguistic discourse context to use as the source for the reconstruction of a sentence (in either the syntactic or semantic sense).

Fifth, it is conceivable that a meaningful language is used by a community that contains only words and phrases but not sentences (Stainton 2006:202–203). In such a scenario, SP is false. Yet the expressions in the language are meaningful. (Moreover, an appeal to Quinean "one-word-sentences" would be beside the point. For then one is using "sentence" in the pragmatic sense of *being the primary unit to perform speech acts* and not in the syntactic and semantic sense at stake.)

For these reasons, I conclude – with Wayne Davis and Robert Stainton – that one should not endorse SP.

Intellectualism Gricean analyses have the problematic consequence that linguistic communication is assumed to be a rational activity performed by rationally deliberating agents which are able to have quite demanding higher-order attitudes.[36]

Millikan (1995a:61–69) has convincingly argued that we should neither assume that communicators have such (implicit) attitudes during ordinary conversations nor that they are able to have such attitudes. Millikan's account (chapter 7) shows that ordinary conversation does not require such attitudes. For it is sufficient that the hearer shows the reaction that is typical for utterances of a certain type. If linguistic communication were only possible by having (and using) the complex Gricean attitudes, then

[36] This is also acknowledged by Grice scholars, *e.g.* Kemmerling (1979, 1986). He works out these features in detail and employs a helpful notation. Schulte (2008:§5.4.3) also argues against Gricean accounts on the basis of the implausibility of such complex attitudes.

linguistic communication would be a complicated Intellectual business. For then, a speaker would need to have complex Gricean intentions when she communicates (by uttering a sentence) with a hearer. And the hearer would need to recognize these complex intentions for communication to be successful. This is an implausible picture of what it is to communicate by using language.

Moreover, young children around the age of three can speak and thereby mean something. But at this age, children do not have second-order beliefs as the second-order false belief task shows; actually, developmental psychology suggests that they do not even possess the required concepts.[37] Plausibly, this holds also for the relevant Gricean intentions.[38] For this reason, children cannot be in the mental states of the type a Gricean analysis would require them to be.

Conclusion The two issues mentioned here have implications for conventionalist accounts. The issue with sub-sentential expressions is problematic for all accounts that entail the sentential primacy thesis. But, maybe more importantly, if an account is Gricean, then it implies an overly Intellectual picture of linguistic communication and unnecessarily assumes complicated attitudes, whose existence is doubtful in the ordinary case.

The situation with respect to Lewis' theory is more complicated. In the chapter on Signaling Games I've argued that Lewis wants his account to be Gricean and that without further rationality assumptions, it does not follow that it is Gricean (§5.3.4 and §5.4). The argumentation there also applies to his Actual Language Relation theory since it's his theory of conventions that needs the rationality assumptions to ensure that the resulting theory is Gricean.[39] Thus, on the one hand, the theory Lewis intended to formulate

[37] This point is also from Millikan (1995a:69), informed by Schulte (2008:§5.4.3). The classic experiment on the false-belief task has been conducted by Wimmer and Perner (1983); their results have been replicated.

[38] Notice that Kemmerling (1986) reconstructs Grice's analysis in terms of beliefs, desires, and instrumental rationality without using intentionality. On his reconstruction, second-order beliefs are used. Thereby, the problematic experimental data on the false belief task directly applies. This is of some importance since Kemmerling's reconstruction is preferable to Grice's analysis: First, it has never been quite clear whether Grice really wanted to state his analysis in terms of intentions and not in terms of beliefs and desires (§5.3.4). Second, Kemmerling's reconstruction makes the rational aspects of communication more transparent and tractable since it fits a decision theoretic explanation.

[39] To get this result, some work is required since we're now using Grice's refined analysis

is Gricean and faces the aforementioned handicaps. On the other hand, one is not forced to make the rationality assumptions. By giving them up, one can defend a version of the theory that is not Gricean (but it would still be committed to the sentential primacy thesis).

6.2.5 Gricean variations

As the evaluation shows, Lewis' theory has problems: Its Gricean commitment makes language use an overly Intellectual activity and comes with an endorsement of the sentential primacy thesis. The conventionality of language turns out to be of a very abstract kind. Instead of words' meanings being conventional, the actual-language relation relating complex entities is conventional. There is no provision for semantic normativity. Applying the theory to linguistically inhomogeneous populations is complicated. Finally, non-literal "Davidsonian" language uses seem to be more problematic than Lewis thought.

There are other conventionalist alternatives to Lewis' Actual Language Relation theory. They've prospered not only because of the appeal of the elementary theory but also because of its modularity. All "modules" come with their set of challenges and different proposals have been made to meet them. In undue brevity, I provide a survey of these proposals with a sole question: Do they improve on Lewis' theory? With respect to the conventional pattern, we can distinguish two types of Actual Language Relation accounts: Gricean accounts and neo-Gricean accounts.

Gricean accounts There are two subtypes, differing in how it is entailed that conventional speakers speaker-mean something: (i) Lewis and Bennett (1973, 1976) developed accounts whose conventional pattern is not stated directly in terms of Gricean speaker-meaning and understanding but together with the definition of a convention, the desired conclusion follows (or should follow; §5.3.4). (ii) Schiffer (1972:156) proposed an account whose conventional pattern is directly stated in terms of Gricean speaker-meaning. Otherwise the explanatory architecture resembles Lewis'. Hence, accounts of either variety do not improve on Lewis' proposal.

of speaker-meaning. On the basis of Kemmerling's reconstruction in terms of beliefs and desires, I'm confident that it can be done.

Moreover, accounts of either subtype provide either no semantic normativity or too much. Without an additional account of social norms, all the *oughts* there are derive from the notion of a convention (see chapter 2). If – as Lewis, Bennett, and Schiffer proposed – the notion of a convention has only a recommending character, then meaning is not normative at all. Otherwise, if the notion of a convention is normative in the relevant sense – as *e.g.* Kemmerling (1976) has it –, then the normativity-of-meaning thesis comes out as a conceptual truth. Both results are undesirable, as I've argued in chapter 2.

Neo-Gricean accounts The second type can be called "neo-Gricean." By this I mean accounts that are not strictly Gricean but still use something like Grice's speaker-meaning. Davis (2003) has developed in detail such an account on the basis of what he calls "cogitative speaker-meaning." Simplifying a bit, cogitative speaker-meaning is defined as the direct expression of an idea (or thought) by producing an utterance (p. 38). Some of the crucial differences are that (i) hearer reactions are not part of cogitative speaker-meaning anymore, (ii) a speaker can also (cogitatively) mean something by uttering a sub-sentential expression, and (iii) a speaker needn't have a belief if she means something; it's sufficient that she entertains a thought.

Davis uses cogitative speaker-meaning together with an account of conventions to define what conventionally used expressions mean (pp. 229–263). There are a couple of problems with Davis' account. As Gauker (2005) has observed, it faces the meaning-without-use problem. In a nutshell, the problem is that Davis' meaning-determination claim assigns meanings to any expression *e* that is derivable from a recursive grammar by means of *e*'s conventional use (Davis 2003:203). But if *e* is too complicated or complex, then it cannot be (conventionally) used.[40]

A second issue concerns his notion of a convention, which is basically a tweaked version of Lewis' analysis of a convention. According to Davis, "[a] convention is a regularity that is socially useful, self-perpetuating, and arbitrary" (p. 206). Conventions are self-perpetuating in virtue of different

[40] There are several problems. Literally understood, Davis' proposal of the conventional pattern makes no sense. For according to his proposal, an agent is said to express a function (Davis 2003:234,257). This can be corrected by taking the values of the functions but thereby, his account has the meaning-without-use problem. Davis acknowledged to me in p.c. that there is this problem.

mechanisms, among them precedent, habit, normative force, social pressure (p. 207). Moreover, "[c]onventions also serve as *generally accepted standards of correctness*. [...] Conventional regularities are thus generally accepted *rules* or *norms*, by which agents judge and guide their actions. [...] [T]hey are de jure rules 'in force' in the community. [...] People use conventional norms not only to guide their own behavior, but also to criticize or correct behavior of others" (p. 211).

With regard to normativity, it seems to me that Davis wants to have the cake and eat it, too. Saying that conventions are rules is not very illuminating since the notion of a rule is as much in need of an explication as the notion of a convention. Moreover, conventions in his sense should be both not demandingly-normative (in case they are sustained by habit) and demandingly-normative (in case normative force and social pressure sustain the convention). In my terminology, some conventions in Davis' sense are conventions with an according social norm in my sense, while other conventions in Davis' sense are just conventions in my sense. So, Davis dodges the conventions/social norms-distinction. But it seems to me that to get a grip on semantic normativity in the context of the conventionalist debate, we should better distinguish between these two notions of "conventions".

6.3 Summary

The basic idea of Lewis' Actual Language Relation theory is quickly told: It relates populations to abstract languages by means of conventions of truthfulness and trust. But not much of it has survived scrutiny and the theory has had to be revised:

(i) The meaning-without-use problem suggests a redefinition of the actual-language relation in terms of $\mathcal{L}$-determining translators (typically realized as grammars).

(ii) Kölbel's lust and lies suggest to modify Lewis' account of conventions by requiring only often-enough conformity to the conventional regularity.

(iii) Non-literal "Davidsonian" language uses necessitate changes of the theory. I vaguely suggested that one could use Recanati's "modulated meaning" proposal to implement these changes. But a good understanding of what happens if one does so is still lacking.

(iv) Common knowledge again gave rise to issues so that it might be best to restrict it or give it up completely.

(v) Due to the Gricean handicaps, it seemed advisable to give up the rationality assumptions.

(vi) The regularities of truthfulness and trust were reinterpreted: from meta-linguistic beliefs about the truth of sentences to the having of certain attitudes with a certain content.

(vii) The extension to polymodal languages (§6.1.4) to deal with other moods creates new problems which seem only to be tractable if we restrict the set of admissible polymodal languages.

As the discussion and evaluation shows, the simplicity of Lewis' theory comes at the cost of the problems summarized above (§6.2.5): it is too Intellectual if Gricean, the conventions are too abstract, there is too little or too much semantic normativity if the *oughts* come from the conventions – or if the *oughts* should come from social norms, then an account of social norms is missing (and we need such an account for more plausible mood explications anyway).

Applying Lewis' cost/benefit-methodology to evaluate philosophical theories (see Hájek 2007), we're invited to consider alternatives. Gricean accounts, I've argued, are unlikely to address these issues. A neo-Gricean account like Davis' seems to be promising. But Davis' account, as developed in (Davis 2003, 2005) has problems, among them the meaning-without-use problem, which might be addressed. I prefer to explore two non-Gricean accounts in the subsequent chapters. Such accounts do not use a Gricean or neo-Gricean notion of speaker-meaning at all. Millikan's account and my alternative account are of this kind.

Let me end this chapter by returning to the conventionality theses C′ and C″. C′ was the version stated in §1. The need to assign meanings to sub-sentential expressions and the meaning-without-use problem led us to reject C′ and to endorse C″:

C′. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

C″. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable linguistic use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

To state C″, the technical term "stable linguistic use" was introduced for linguistic conventions and linguistic social norms (and combinations of them), where such conventions and social norms are ones whose dispositions are in part brought about by a translator.

Chapter 7

Towards an Evolutionary Theory

> [T]he best classification of the various languages would [...] be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue.
>
> *The origin of species*
> Charles Darwin

Already Charles Darwin had observed an analogy between biological evolution and the evolution of languages. But only two centuries later was an account of language developed that takes this insight to heart: In 1984, Ruth Millikan presented in her book *Language, thought, and other biological categories* (Millikan 1995a) a novel evolutionary account of language.[1] Her account is one of the accounts discussed in this chapter. The other account is Simon Huttegger's which belongs to the family of Evolutionary Signaling Games.

I begin by discussing Millikan's account in §7.1. In §7.2 I discuss Evolutionary Signaling Games. Both have open problems. This leads to an outlook for a better theory in §7.3 and a summary in §7.4.

7.1 Millikan's account

On Millikan's conception, linguistic behavior can be portrayed as consisting in solutions to long-term signaling problems. But in contrast to Lewis'

[1] Millikan elaborated on her account subsequently in a series of lectures and papers, published as *Varieties of meaning* (Millikan 2006) and *Language – a biological model* (Millikan 2005b), respectively.

rationalistic conception, it's not only rational deliberation that can bring about coordination but also *evolutionary adapted behaviors* (in the wide sense of "evolution" including cultural evolution[2]): Suppose the language users have acquired a tendency to bring about outcomes that are good enough for their purposes often enough. They could have acquired such a tendency through a learning process like imitation. Then having this tendency is sufficient for there to be a convention in Millikan's sense. So, one important difference between Lewis' account and hers is that his account of conventions is rationalistic while hers is evolutionary.

But the differences go deeper. Millikan defines key notions such as *expression*, *meaning*, *language*, and *convention* in original non-standard ways. For example, expressions are individuated by the use histories of their tokens. This differs from the standard understanding in linguistics and philosophy according to which expressions are individuated by their physical form (sound, shape).[3] Common to all these notions is that they are defined in terms of biological purposes ("proper functions"). In virtue of these, expressions mean what they do. Roughly speaking, an expression's meaning is explained in terms of the proper function its conventional use has.

7.1.1 Proper functions

Millikan developed a theory of proper functions in the first two chapters of her book *Language, thought, and other biological categories* (Millikan 1995a). Informally, the theory is meant to make claims about *purposes* or *functions* precise – for example when one claims that "the function of the heart is to pump blood." The goal of the theory is to provide a naturalistic explanation of mental content and linguistic meaning in terms of proper functions. Let me sketch the big picture as it applies to her project about language before I present her definitions.

According to Millikan, we should conceive of language as a "tangled jungle of overlapping, crisscrossing traditional patterns."[4] She often calls these linguistic patterns *language devices.* They are characterized as "any significant linguistic surface element that a natural spoken or written language may contain like words, tonal inflections, stress patterns," including

[2] Richerson and Boyd (2006) provide a state-of-the-art exposition of such a theory.

[3] A similar proposal has been made by Kaplan (1990).

[4] See (Millikan 1998:176).

the grammatical moods like indicatives.[5] The analogy between linguistic items and devices is meant quite strictly as grammatical moods illustrate: they are things that are good for something.

Indicatives, for example, have the proper function *to produce or activate true beliefs in hearers.*[6] This explains what the device "indicative mood" is *for.* For this reason, a device's proper function can also be understood as its survival value: it explains why it's kept in currency.[7]

That a language device can perform its proper function requires that certain "Normal conditions" are satisfied.[8] In case of indicatives these are that speakers are reliable truth-tellers and that the hearers' sign-consuming mechanisms are working (*i.e.*, produce or activate the relevant belief when an indicative is recognized).[9] Normal conditions needn't *always* be satisfied but just often enough. Otherwise, it would, at some point, not be beneficial anymore for speakers or hearers to do their part in the speech act pattern.

So, language devices are attributed a proper function to explain what the device was historically for and this also explains why the device proliferates. To this end, Millikan defines "proper function" roughly as follows (*cf.* p. 26):

PF. A thing x has the proper function to ϕ iff x exists and is the way it is because the ancestors of x have performed ϕ sufficiently often.

This a simplified version. It will do for her linguistic theory. The definition makes use of other notions: "Ancestor" is defined relative to so-called "reproductively established families." A *reproductively established family* is a set of similar things which are similar because they have been reproduced in the same way. The relevant ("first-order") reproductively established families of our concern are sets of things which have been reproduced from the same original (or "model"; p. 23).

A thing c ("copy") is a *reproduction of* another thing o ("original") iff there is a causal mechanism to produce o-like things based on the model

[5] See (Millikan 1995a:3).

[6] See (Millikan 1995a:§3).

[7] See (Millikan 2005e:92 ff.).

[8] The capitalized "N" in "Normal" indicates that the word is used in a stipulated sense which is not the same as *on average* or "often" (Millikan 1995a:5).

[9] Millikan does not state the Normal conditions for indicatives explicitly but see (Millikan 1995a:53,58) for her characterization of their proper function and (Millikan 2006:76 ff.) for an illustration of her account of representational content in terms of sign-producers and -consumers. It seems that the proper function can only be performed if the conditions I suggest are satisfied.

of o such that if o had been different in specifiable respects, then its copies would have differed accordingly, and c has been produced by this mechanism (pp. 19–23).

A thing x is an *ancestor of* a member y of a reproductively established family if x occurs earlier in a reproduction chain which results in y (p. 27 f.). (Only "if" and not "iff" because there are other ways to be an ancestor that I've omitted for simplicity.)

Ancestors of something having a proper function ϕ didn't have to perform it frequently but at least so often that it was still beneficial to be reproduced (p. 29). The relevant sense of "beneficiality" here is the one from evolutionary theory that explains a thing's survival in an environment.

Let me provide some examples to illustrate the definitions: (i) It is true to say that a heart has the proper function to pump blood (ϕ) iff the heart exists and is the way it is because its ancestors have pumped blood sufficiently often.[10] (ii) Imitation is a reproduction mechanism (p. 21). (iii) Handshaking has the proper function to greet iff handshakes exist and are a means to greet because by shaking their hands, people greeted each other in the past sufficiently often. Plausibly, a behavior like this has been reproduced by imitation.

Linguistic expressions are another example. According to Millikan, "[w]hat makes two word tokens tokens of the same word [. . .] is a matter of the history of these tokens. Word tokens are classified into types by reference to reproductively established families [. . .]" (p. 75):

MI1. Expression types are reproductively established families of expression tokens. That is, expressions are individuated by the reproduction chain of patterns in which the expressions occur.

Consequently, the word spelled as *r-o-t* ("rot") in (what is commonly called) *English* is not the same word as the word spelled as *r-o-t* in (what is commonly called) *German* since, presumably, they come from different lineages. If tokens of the so-spelled words had the same history, then they would be tokens of the same word.

[10] Strictly speaking, for such cases we have to re-define the terms "ancestor" and "reproductively established family" in a way that is more general. For hearts do not reproduce hearts. Rather, genes are reproduced and they bring about hearts. Millikan provides these definitions in (Millikan 1995a). For my presentation, I think it's not worth to go into these details. For the proper functions of language devices like words, one should keep in mind that there are similar complications.

Teleological *oughts* Another feature of proper functions is that they determine something like a *standard of comparison*: Actual performances of a thing having a proper function can be compared as to whether they accord with the proper function or not. This is not to say that proper functions have a demandingly-normative character:

> The task of the theory of proper functions is to define this sense of "designed to" or "supposed to" in naturalist, nonnormative, and nonmysterious terms. (Millikan 1995a:17)

Millikan's explication of what "supposed to" means is in terms of proper functions: x is supposed to do ϕ iff x has the proper function to ϕ. We can express this also as "x ought to ϕ" but we must keep in mind that such *oughts* are explained in terms of the thing's history; in particular, these *oughts* don't have a demanding character. Nevertheless, for an explanation of the normativity of mental content such *oughts* might be sufficient, as Millikan proposes:[11] the paradigm is a belief that *ought* to be true in order to perform its function. But I won't delve into this topic, and instead will focus on semantic normativity.

Millikan (1990) proposed that the *oughts* definable in terms of proper functions are constitutive for semantic normativity. But on the basis of my argumentation in chapter 2, in particular §2.2.3 on semantic mistakes, we have good reason to reject this proposal: *Just because* a language device has a certain proper function, a person cannot be demanded to use it in accordance with its proper function. The reason is that proper functions are non-evaluative and lack a demanding character.[12] Hence, on the basis of proper functions, semantic normativity cannot be explained as required.

7.1.2 Meaning

Millikan's position with respect to meaning is original. This starts with meaning assignments. Since Millikan individuates expressions more finely than usual on the basis of their histories (MI1), reference to a language is not required anymore:

MI2. Meaningful expressions don't have a meaning relative to a language.

[11] See (Millikan 1995a:1).

[12] Millikan agrees on that point, see (Millikan 2005b:v. ff.) and (Millikan 2008).

MI1 together with MI2 have the additional consequence that it's not *at all* required to use a notion of a language to state her account.[13] Thereby, a fruitful perspective to explain language varieties is opened. I return to this in §7.3.1 and §9.4.

Millikan's account is also original in that language devices, including expressions, usually have meanings of *two* different kinds,[14] namely their proper functions (called "stabilizing proper functions") and so-called "semantic mapping functions", which are akin to but not identical to satisfaction conditions.

Stabilizing functions *Stabilizing functions* are so called because they explain what keeps language devices in currency. The stabilizing function of a language device is one "that, when performed, tends both to encourage speakers to keep using the device and hearers to keep responding to it with the same (with a stable) response."[15] Since stabilizing functions are a special case of proper functions, we already know under which conditions a language device is assigned a particular stabilizing function.[16]

For the central moods *indicatives*, *imperatives*, and *interrogatives*, Millikan stipulates the following stabilizing functions: Indicatives have the stabilizing function of producing or activating true beliefs in the hearer. Imperatives have the stabilizing function of producing hearer compliance. Interrogatives have the stabilizing function of eliciting true answers.[17]

Stabilizing functions of moods are not too different from mood explica-

[13] This is surprising since Millikan's article *In defense of public language* (Millikan 2005a) suggests that there is a role for languages. Millikan confirmed to me in p.c. that both observations – (i) a notion of language is not required while (ii) the mentioned article suggests to the contrary – are correct.

[14] There is a third kind of meaning, the private *conceptions* individuals associate with a language device. I ignore them since they are private and are not used to explain in virtue of what meaningful expressions have their public meanings.

[15] See (Millikan 2005d:94).

[16] More precisely, the stabilizing function of a language device is the *focused* proper function it performs in a variety of performances under Normal conditions (Millikan 1995a:34–38). While a device can have many proper functions, its focused proper function has a special status, based on causal relationships between the various proper functions; keeping the body alive is a proper function of the heart, as is providing oxygen to the muscles, but its focused proper function – causally prior to both of these – is pumping the blood.

[17] These stipulations are meant as empirical hypotheses. Millikan has made this proposal in several writings, *e.g.*: (Millikan 1995a:§3), (Millikan 2005c:63), (Millikan 2005d:94), (Millikan 2006:25–27).

tions known from other theories such as Lewis'. But there is one important exception that renders Millikan's account non-Gricean:[18] if a speaker performs a speech act, then she does not have to *intend* that its stabilizing function is performed (and likewise for the hearer: she does not have to *recognize* a speaker's intention). In case of indicatives, this is to say that it suffices that a speaker utters a sentence that is true and that the hearer comes to believe what the sentence says.

Not only moods are supposed to have stabilizing functions but also meaningful expressions. If I understand the works of Millikan correctly, then she has no elaborate theory of the stabilizing functions of expressions of common syntactic categories. For sentences the idea is, however, reasonably clear. Take for example the sentence "It is raining." Plausibly, the sentence in its indicative mood has the stabilizing function of producing or activating the true belief in the hearer *that it is raining.* A further plausible hypothesis is that the sentence in its indicative mood has this stabilizing function because utterances of the sentence correlated in the past in systematic ways with raining-facts, *i.e.* facts of the sort that it has rained at some place for some time. So, it seems that the stabilizing function of the *sentence* "It is raining" is one in relation to or in support of the sentence in its indicative mood.

Millikan endorses the sentential primacy thesis.[19] Hence the stabilizing functions of words and other sub-sentential expressions are derived from the stabilizing functions of sentences: They have whatever stabilizing function allows the sentences to perform their stabilizing functions. A consequence of this is that the stabilizing functions of sub-sentential expressions are underdetermined (see the discussion of the sentential primacy thesis in §6.2.4).

Semantic mapping functions as meanings Linguistic expressions are intentional (in the *aboutness* sense). To capture this feature of linguistic expressions, *semantic mapping functions* are assigned to language devices. Many but not all expressions that have a stabilizing function also have a semantic mapping function. A semantic mapping function maps expression tokens to worldly entities (objects, properties, relations, states of affairs).[20]

[18] See (Millikan 1995a:52 ff.).

[19] See (Millikan 1995a:80, 104).

[20] *Cf.* (Millikan 1995a:9 ff.).

An example of a meaningful expression that has a stabilizing function but not a semantic mapping function is “Hello!”:[21] It is used as a means (“device”) for greeting but it does not represent the world.

There are many semantic mapping functions, assigning different values to expression tokens. If an expression has a semantic mapping function, then it’s determined by its stabilizing function:

> [T]he [stabilizing function] of a sentence is something which, having itself been made determinate, in turn determines the conditions under which the sentence is true. [...] What determines the meaning of a sentence is then what determines the mapping functions in accordance with which it must map onto something in the world in order to be true. Put roughly, the meaning of a sentence is its own special mapping functions—those in accordance with which it “should” or “is supposed to” map onto the world. (Millikan 1995a:9)

For example, tokens of the sentence “It is raining” correlate, under Normal conditions, in systematic ways with certain raining-facts. This correlation between indicative tokenings of the sentence and the state of affairs can be described by a semantic mapping function. It is necessary that such a correlation exists for the sentence’s having the stabilizing function it has.[22] Hence, we can state Millikan’s thesis about this kind of meaning as follows:

MI3. Expression e means semantic mapping function $\mathcal{M}$ iff (i) e has the stabilizing function ϕ and (ii) for all tokenings t of e: if t performs ϕ, then t maps onto world affairs according to $\mathcal{M}$.

Semantic mapping functions resemble both satisfaction conditions and interpretation functions but there are important differences. First, an expression’s meaning in the sense of semantic mapping functions is *not* a satisfaction condition but a function that assigns a satisfaction condition to *tokens* of the expression. Hence, we could say that expressions are assigned satisfaction conditions relative to utterance situations by their semantic mapping functions.

Second, in contrast to interpretation functions that assign denotations to all expressions of a language, a semantic mapping function assigns values only to tokens of *some* expressions. Each expression that has a semantic mapping function has its own. Thus, there is not *one* semantic mapping

[21] See (Millikan 2005d).

[22] Collins (2009:162-166) also understands Millikan’s proposal in this way.

function for *all* expressions. Third, a token of a meaningful expression only has a value assigned by its semantic mapping function if the Normal conditions of its stabilizing function are satisfied. For example, the Normal conditions for a proper name include being part of a sentence and there being something onto which it historically mapped.[23] Thus, a name outside the context of a sentence has no value at all and neither do empty names like "Santa Claus." Fourth, a semantic mapping function of an expression is not only defined for tokens of the expression but also for tokens of *alternative* expressions that have the same syntactic form:

> The semantic-mapping function is given by rules according to which significant transformations of the sentence that conserve its syntactic form yield different truth- or satisfaction-conditions. Compare the sentence "It's raining" with the sentence "Rain is falling here now". "It's raining" contrasts with "It's snowing", "It's hailing", "It's sleeting", and so forth. All display the same syntactic form, the transformations that substitute "snow", "hail", and "sleet" for "rain" determining different satisfaction-conditions in a systematic way. Similarly, "Rain is falling here now" contrasts with "Snow is falling here now", "Hail is falling here now", "Sleet is falling here now", and so forth, but it contrasts, further, with "Mist is rising here now", and with "Rain was falling in Rome yesterday". The truth-conditions of "It's raining" and of "Rain is falling here now" are the same, but the semantic mapping is different. "Many drops of water are presently precipitating from the atmosphere and landing close to this place" also has the same truth-condition but is articulated by yet another semantic-mapping function. (Millikan 2005c:63–64)

Millikan's idea is that linguistic expressions (and intentional signs in general) must come in "articulated" systems such that their transformations (according to some syntactic operations) relate to corresponding transformations of their semantic values (according to some semantic operations).[24]

[23] See (Millikan 1995a:105 ff.).

[24] In her recent writings, Millikan often claims that semantic mapping functions are *isomorphisms* between the linguistic system and the world; see (Millikan 2006:49) and (Millikan 2005e:97). But this is implausible since it would rule out synonymy. Moreover, the claim can only be understood in a trivial sense by means of setting up the domain in way that there is a name for every object in it, as isomorphisms require. If the claim were to be understood in a non-trivial way, then plausibly this requirement wouldn't be satisfied. These issues can be avoided if we require the function to be a *homomorphism.* The problem has also been observed by Shea (forthcoming), suggesting the same solution to it. But with-

7.1.3 Conventions

Millikan's notion of a convention is motivated by the difficulties Lewis' account faces (Millikan 1998). To remind you of a few: It turned out to be problematic to assume that conventions ensue near-perfect regularities of conforming behavior. Common knowledge is required but hard to get. Conventional behavior seems to be a matter of social habits and not of rational deliberation. As we've seen in §4.3, some of the issues can be addressed. But the resulting account lost some of its appeal.[25]

Millikan proposed new definitions:[26]

A pattern of activity R is called a "natural convention" iff (i) R has been reproduced for a while and (ii) R proliferates partly because of weight of precedent.

A natural convention is a "coordinating convention" iff its pattern of activity R satisfies the following properties:

1. R involves more than one participant.
2. R is proliferated because its instances serve a purpose had in common by the participants.
3. R is such that the contribution to the joint pattern that each participant must make in order to reach the common goal depends crucially upon the contribution made by the other(s).
4. R is such that a variety of equally viable alternative joint patterns would achieve the same goal as well.

Conventions in Millikan's sense have a couple of important features. First, *natural conventions* are not conventions according to my pre-theoretic characterization (§1.2) since coordination is not necessary. To button shirts in certain ways can count as a convention in this sense.[27] In contrast, coordination conventions *are* conventions in the pre-theoretic sense since the conditions their patterns have to satisfy imply that (i) there is a Lewisian coordination problem among the participants that is solved if they do their parts in it and that (ii) the parties to such a convention are cooperative.

Second, coordination conventions are *dispositional conventions* (§1.2) since conforming to such a pattern of activity often enough is sufficient for there to be a coordination convention. This can be brought about by the agents' having suitable dispositions. The agents need not be in certain

out defining the algebras on which the allegedly isomorphic semantic mapping function is defined, these claims are of little interest since they have no determinate content.

[25] My changes to Lewis' account were inspired by Millikan's account.

[26] The first is from (Millikan 1998:162) and the second is from (Millikan 2005a:40).

[27] See (Millikan 1998:163).

epistemic states; *e.g.* they need not believe or know that the others conform. Since imitation also counts as a reproduction mechanism, such mechanisms can be non-Intellectualist.

Third, the existence of a coordination convention does not imply that there is a near-perfect convention-conforming regularity. The conventional pattern has only to be conformed to often enough.

Fourth, Millikan conventions are not defined relative to a group or an environment. Conventional patterns are, however, reproduced in an environment by certain agents. If we like, we can make this explicit: The *group* of a conventional pattern R is the set of agents that have performed the conventional pattern R in the past and have a disposition to act conformingly. The *environment* of a conventional pattern R is the environment in which R has been exhibited.

7.1.4 Recovering the conventionalist account

The central theses MI1–3 are not stated in terms of conventions. Nevertheless, Millikan's account is conventionalist. The key is the correspondences between conventions and proper functions: Natural conventions correspond to a class of reproductively established families of patterns of activity that proliferate partly because of weight of precedent. Coordination conventions correspond to a subclass of this class, namely the one whose patterns satisfy the four conditions of coordination conventions. (The "purpose" mentioned in the second condition should be understood in the sense of *proper function.*) In case of linguistic patterns, the proper functions should, I think, be their *stabilizing functions* since they explain why the patterns continue to exist as they are.[28] This correspondence helps us to reconstruct Millikan's conventionalist account as follows:

MI4. Expression e means semantic mapping function $\mathcal{M}$ iff e's use pattern is a coordination convention which proliferates because (i) e has the stabilizing function ϕ and (ii) for all tokenings t of e: if t performs ϕ, then t maps onto world affairs according to $\mathcal{M}$.

7.1.5 Evaluation

Millikan's account is an important contribution to the conventionalist project. It allows to explain rational behavior without making rationality

[28] *Cf.* (Millikan 1995a:31 ff.).

assumptions. This has the welcome consequence that her account doesn't face the Intellectuality charge, a central problem of Gricean accounts. The two reasons for this are: (i) it uses an evolutionary notion of a convention and (ii) the communicative pattern is described in non-Gricean terms. So, except for the problems resulting from the sentential primacy thesis (see below), Millikan's account doesn't face the handicaps of Gricean accounts (§6.2.4). Moreover, Millikan forced us to rethink common conceptions thereby advancing our conceptual understanding. Yet, her account has problems which influence my subsequent evaluation on the basis of the adequacy conditions of §1.4.

From stabilizing functions to semantic mapping functions The crucial elements in Millikan's account are *stabilizing functions.* The conventional uses of expressions in accordance with them determine their meanings. Millikan tells us that the use in accordance with a certain semantic mapping function is a necessary condition for the use of the expression serving its stabilizing function. Thus, by knowing an expression's stabilizing function (and its Normal conditions), one should be able to derive its semantic mapping function *somehow.* But Millikan does not say how one can derive these semantic mapping functions. Hence, Millikan's theory is vague and incomplete with respect to this step. One way to close this gap is define a class of stabilizing functions that are structured in such a way that semantic mapping functions are one of their constituents.[29]

Syntactic structure and a looming holism Another area where Millikan's account is underdeveloped is *syntactic structure.* Millikan sometimes talks of a sentence's "syntactic form" (*e.g.* Millikan 1995a:53) and often alludes to a Chomskian Universal Grammar. For example, in her article *In defense of public language* (Millikan 2005a:36), Millikan seems to suggest that an expression's syntactic form is determined by the grammars the language users possess.[30] However, Millikan does not elaborate on the rela-

[29] Millikan probably wouldn't accept this proposal. For semantic mapping functions encode information about the *Normal conditions* of the expression's stabilizing function. Thus, they would be part of a stabilizing function. But Millikan wants to separate them from the function: thereby she can explain why having a function in a different environment is not adaptive anymore.

[30] Millikan (2008) suggests that a construction-grammar approach fits Lewis' account better. The points I make also apply to construction grammars.

tions between language devices, their stabilizing functions, and grammars. In particular, she does not define the syntactic form of sentences and the sub-sentential expressions they are made up of. This has three unwelcome consequences.

First, if one endorses the sentential primacy thesis, then one is committed to say how the meanings of sub-sentential expressions are to be derived. Otherwise, there is an explanatory gap with respect to the meanings of sub-sentential expressions.

Second, the stabilizing function of a sub-sentential expression plausibly consists, at least in part, in *being a part of sentences in systematic syntactic environments thereby systematically contributing to the stabilizing functions of these sentences*.[31] Without defining syntactic forms, such a description of a function has no determinate content. Since the meaning-determination claim MI4 is stated in terms of stabilizing functions, it has also no determinate content.

Third, her meaning-determination claim is, on plausible assumptions, holistic in a problematic way so that thesis H follows if we understand "meaning" in the sense of *semantic mapping function*:

H. For any language $\mathcal{L}$: The meaning of (almost) any expression of $\mathcal{L}$ depends on the meanings of (almost) all other expressions of $\mathcal{L}$; *i.e.* each difference between $\mathcal{L}$ and another language $\mathcal{L}'$ implies that the expressions in $\mathcal{L}$ have other meanings than the expressions in $\mathcal{L}'$.

An example illustrates my claim: Let $\mathcal{L}$ be the language consisting of the words "Antonius", "Brutus", "Caesar" (the Ns), and "lives". Sentences of $\mathcal{L}$ have the form "N lives" wherein an N may take the position of "N"; *e.g.* "Caesar lives" is a sentence of $\mathcal{L}$. Suppose further (without loss of generality) that the semantic mapping functions are homophonic. Then the semantic mapping function $\mathcal{M}_B$ of "Brutus" is such that $\mathcal{M}_B$ assigns "Brutus" the value *Brutus* in the context of sentences of $\mathcal{L}$.

Since semantic mapping functions have the transformation property, a semantic mapping function $\mathcal{M}_N$ of an N is also defined for all other Ns (N's alternatives). Hence, $\mathcal{M}_B$ also assigns values to "Antonius" and "Caesar". Plausibly, the transformation property is such that (1) is true:

(1) For all Ns n and n' (with $n \neq n'$): for all semantic mapping functions $\mathcal{M}_n$ of n and $\mathcal{M}_{n'}$ of n': $\mathcal{M}_n(n') = \mathcal{M}_{n'}(n')$.

[31] Millikan seems to hint at this in (Millikan 1995a:107).

According to (1), a semantic mapping function of a certain N also assigns homophonic values to N's alternatives. Requiring that semantic mapping functions have the transformation property but rejecting (1) seems to me to be incoherent. For this reason, the transformation property should be understood so as to imply (1).

Therefore: If "Caesar" had another semantic mapping function, then the semantic mapping function of "Brutus" would be different. From this it does *not* follow that the respective semantic mapping function $\mathcal{M}_B$ would assign a different value to "Brutus". It only follows that $\mathcal{M}_B$ assigns "Caesar" a different value. Yet, since semantic mapping functions *are* meanings, we can put the consequence as follows: If a name in $\mathcal{L}$ meant something else, then all the other names of $\mathcal{L}$ would have different meanings. Hence, the problematic thesis H is true in this case.[32]

The case can be generalized since semantic mapping functions $\mathcal{M}$ have the transformation-property: The semantic mapping function $\mathcal{M}_e$ of a particular expression e is also defined for tokens of *alternative* expressions e' *having the same syntactic form.* Plausibly, any two noun phrases count as having the same syntactic form. If so, their semantic mapping functions must be defined for all of them. Since a noun phrase can embed almost all other expressions (by means of relative clause constructions), a noun phrase's semantic mapping function must be defined for almost all expressions. Hence, H is also true for complexer languages.

But surely, an expression's meaning shouldn't necessarily depend on almost all expressions.[33] Even Millikan thinks that such a holism would be "disastrous."[34] The problematic consequence can be avoided by giving up the transformation property.

For these reasons, I conclude that Millikan's account can only be defended if one defines the syntactic forms of expressions in much more detail.

Adequacy of Millikan's account of conventions Millikan's account of dispositional conventions is adequate (on the basis of the adequacy con-

[32] As Peter Schulte pointed out to me in p.c., Millikan could reply: "But the meaning of 'Brutus' doesn't change!" – It's correct that the semantic value $\mathcal{M}_e(e)$ for e does not change, if the meanings of other expressions e' change. But the semantic mapping function $\mathcal{M}_e$ would change and that *is* the meaning.

[33] Holism is discussed in general in (Fodor and Lepore 1992); Horwich (2004) discusses holism in connection to use theories of meaning.

[34] See (Millikan 2000:100).

ditions from §1.4). It would be interesting to develop her account formally. In its present form, it's hard to derive empirically testable predictions. This is not a substantial objection but is merely meant to point out *what to do next*. The framework of evolutionary (game) theory offers many precise (and potentially testable) models.[35] (I'll briefly return to this topic below.) For this reason, I think the outlook of her account is promising.

Adequacy of Millikan's account of social norms Millikan's account of social norms derives from her account of proper functions. Since it is unclear how to explain the demanding character of social norms in this sense (DesN1), her account is incomplete.

Adequacy of Millikan's account of meaning A strength of Millikan's conception of an expression is that meanings can be attributed to them without having to refer to a language. As we will see, this allows us to make progress in the explanation of language varieties; see §9.4.

The problems in the normativity-department have repercussions for Millikan's account of meaning: The normativity resulting from an expression's having a stabilizing function seems to be something else than semantic normativity. Making a semantic mistake is not *just* doing something that is not in accordance with an expression's stabilizing functions (DesM3). It's doing something one ought not to and this ought has a demanding character. The demanding character isn't explained by appealing to facts about the stabilizing function of an expression. Moreover, if normativity in Millikan's *proper-function* sense is used to explain semantic normativity, then it follows that necessarily, meaning is normative. I've argued in chapter 2 that this is wrong. If semantic normativity should be explained differently, then Millikan's account would be incomplete since the explanation is not provided.[36]

By endorsing the sentential primacy thesis, Millikan faces the difficulties resulting from that, *e.g.* the problem of deriving sub-sentential meanings and their indeterminacy (§6.2.4 for other issues).

The use of conventions to explain an expression's meaning yields an implausible solution to Humpty Dumpty's problem, namely that it always de-

[35] A recent review of models of language evolution is (Boer and Zuidema 2009).

[36] The alternative also goes against the MD-ME principle promoted in chapter 2: if meaning is normative, then this is because of the way it is determined.

pends on how every member of the convention uses the expression (DesM5).

Finally, her account is problematic with respect to syntactic structure (DesM6).

7.2 Evolutionary Signaling Games

Evolutionary Signaling Games accounts improve on Lewis' classical Signaling Games theory by reinterpreting the formalism in an evolutionary way. Thereby, the Intellectuality charge can be addressed. In this section, (i) I indicate how Millikan's account resembles Evolutionary Signaling Games accounts and I argue that there are comparative benefits to using the the latter instead of Millikan's account. (iii) I briefly discuss the theory of Huttegger (2007) as an example of such an evolutionary theory that in addition adds a basic account of social norms. I argue that if one is inclined to endorse Millikan's account, one should rather endorse an Evolutionary Signaling Games account like Huttegger's. (iv) I return to the *no-structure* problem of signaling games. Recent research has improved on it. Thereby, the outlook of (Evolutionary) Signaling Games becomes even more promising.

7.2.1 Millikan's account as an evolutionary signaling game

Let us focus on indicatives to establish the similarities. According to Millikan, an indicative is used in accordance with its stabilizing function iff the hearer comes to believe truly what the uttered sentence says. This is only so if the function's Normal conditions are satisfied: the uttered sentence must be true. The use pattern "sentence is true – speaker utters the sentence – hearer believes what it says" is similar to the use pattern at which we arrived when discussing Lewis' Signaling Games theory (chapter 5). According to Lewis, the speaker has to believe what the sentence says and the hearer needs to believe that – whether the belief is true or not. According to Millikan, the speaker does *not* need to believe what the sentence says and the hearer's belief must be *true*. But we shouldn't overstate the differences. Plausibly, a speaker Normally speaks truly because she believes what she says. Moreover Schulte (2008:177–179) has argued that the descriptions of proper functions have to be hedged by *ceteris paribus*-assumptions: Indicatives can only perform their stabilizing functions, if the belief-systems of speakers and hearers are functioning Normally. If we add the *ceteris paribus*-assumption and claim that the stabilizing function of indicatives is *to produce or activate beliefs in hearers*, then it follows, *ceteris paribus*, that the beliefs a hearer acquires by hearing indicatives will Normally be true.

With identical use patterns, the main difference between Millikan's account and Lewis' Signaling Games theory consists in the notion of a convention.

From Millikan's account to Evolutionary Signaling Games Millikan's notion of a convention is not the only evolutionary one. Such notions (called "solution concepts") are studied in the field of evolutionary game theory. Common to such theories is that they reinterpret the game theoretic formalism in a way that its agents are not required to deliberate about what to do. It suffices that agents have evolutionary adaptive behavioral dispositions to behave in certain ways which are sensitive to what the others do (in §1.2 I called such dispositions "effective coordinative dispositions").

An important notion from theoretical biology is the notion of an *evolutionary stable strategy* (ESS) which captures the idea that the agents' evolved strategies are resistant against changes (if there were a change, it wouldn't spread in the population).[37] In terms of ESSs, Robert Sugden defined conventions as evolutionary stable strategies in coordination games (see Sugden 1986:32).

The notion of an ESS has also been applied to signaling games. A general result is that under certain weak assumptions, efficient communication systems evolve.[38] Such an explanation of signaling is attractive since it does not rest on the problematic rationality assumptions of Lewis' Signaling Games theory.[39]

Moreover, evolutionary game theory has comparative benefits over Millikan's account. For one thing, there are precise relations between a popular evolutionary dynamics (the replicator dynamics) and learning theory.[40]

[37] *Cf.* (Maynard Smith and Price 1973). A strategy S is *evolutionary stable* iff for all other strategies T: (i) $u(S,S) > u(S,T)$ or (ii) if $u(S,S) = u(S,T)$, then $u(S,T) > u(T,T)$.

[38] For example, Brian Skyrms (1994, 1998) uses evolutionary stable strategies as his solution concept with the important additional assumption that the agents' strategies are correlated.

[39] As in classical game theory, there are foundational issues in evolutionary game theory, *e.g.* relating to the notion of utility. See (Alexander 2009:§5) for an overview.

[40] See for example (Lenaerts and Vylder 2005). Tom Lenaerts and Bart de Vylder study a

Millikan does not have such a foundation and her appeal to imitation learning is vague. Also, there are different theory-internal explanations for the agents' cooperation whereas Millikan has to assume it.[41]

7.2.2 Huttegger's theory

Of particular interest is the proposal of Simon Huttegger (2007), building on the work of William Harms (2004:§8) – which was inspired by Millikan's. For reasons that will become clear as we go on, Huttegger's theory can be understood as a partial implementation of Millikan's account in the framework of Evolutionary Signaling Games. Moreover, Huttegger innovated on standard Signaling Games accounts by adding a basic account of social norms.

Harms' idea was to understand social norms as a kind of special signaling game whose types (the states the world can be in) are states of non-conformity to certain behaviors which have evolved in a certain population. The signals of such signaling games are assigned a complex meaning which Harm's calls "primitive content" corresponding to the two functions they have: (i) they indicate that someone did not comply with a "rule" and (ii) they demand to behave conformingly (Huttegger 2007:273). Such meanings can be used as meanings for imperatives, as Huttegger proposed. The *demanding*-function implements the characteristic sanction mechanism of social norms. Harms' proposal is then to identify social norms with evolved behaviors that come with such a sanction mechanism.

Huttegger formalized Harms' idea in terms of what he calls a *normative system*. Roughly speaking, a normative system is a pair of games consisting of a basic underlying game and a superimposed signaling game such that (i) there is a regularity to conform to one of the strategy profiles of the underlying game in some population, (ii) this regularity has evolved,

model of learning meaning-word associations. They assume a "language acquisition process wherein the listener is not observing some communication but actively participating. The importance of this participation is that the hearer/listener can become speaker in the next round [...] From a cultural learning perspective this model belongs to the class of operant-condition models" (p. 264), to which also reinforcement learning belongs. Hofbauer and Sigmund (2003) provide a recent survey of evolutionary game dynamics, also relating it to learning theories; in particular, see p. 504 for references about models of reinforcement learning that are closely related to the replicator equation.

[41] *E.g.* Axelrod's repeated tit-for-tat (Axelrod 2006) and Trivers' theory of reciprocal altruism (Trivers 1971).

and (iii) the "superimposed" signaling game is of Harm's kind – that is, it furthers conformity to the regularity in the population. Conditions (i) and (ii) can be understood as requiring the strategies to be evolutionary stable strategies.

Now, let me return to my claim that Huttegger's proposal can be understood as a partial implementation of Millikan's account. I outlined above how Millikan's account can be understood as a Signaling Games account using an evolutionary notion of a convention. This relates as follows to Huttegger. A special case of Huttegger's general theory is a normative system whose underlying game is a signaling game; we can call such a normative system a "normative signaling system." The underlying signaling game in a normative signaling system can be understood as an explanation of the teleological *oughts* that Millikan explains in terms of proper functions. For Millikan's proper functions correspond to signaling in accordance with the evolutionary stable strategy of the underlying signaling game which prevails among the members of the normative signaling system.

The superimposed game of a normative signaling system goes beyond Millikan's proposal. It can be understood as a basic account of social norms. For deviations from the evolutionary stable strategy (how one "ought" to behave) are punished. That is to say, if the signals of the underlying game are not used in accordance with their meanings, then there is a punishment. Arguably, if one is able to punish someone, then one is in a position to demand something. Hence, it's only a short step to say that uses of the signals in the superimposed game have a demanding character. Since the punishments can be the result of simple reactive mechanisms of non-Intellectual agents, we come close to an evolutionary non-Intellectual explanation of semantic normativity on the basis of the Harms/Huttegger proposal: The members of a population in which a normative signaling system exists can demand that the others use the signals in accordance with their meanings.

7.2.3 Signaling games with linguistic structure

Signaling Games are often criticized for having unstructured signals and only finitely many of them. Hence, the objection goes, one cannot explain the combinatorial features of language and assign meanings directly to words. This criticism is too quick for two reasons.

First, recent developments by Martin Nowak *et al.* have improved on the problem of structure. In a series of publications,[42] they have developed different evolutionary models that use structured signals and structured meanings. The structure they consider is of the kind "Name Verb." While simplistic, it is theoretically interesting. For such models can explain recombinations of parts, for example the recombinations resulting from the use of "Peter drinks" and "Mary lives" to "Mary drinks" and "Peter lives." Thereby, these models can be said to use a notion of a word and can attribute to them meanings in a way that avoids the problems resulting from endorsing the sentential primacy thesis (§6.2.4). However, the models restrict complex signals to finite recombinations of finitely many parts, hence yielding a finite language with no recursive structure. Recursive rules are not studied. That is to say that there is nothing like embedding or an "and" in the syntax.

Second, there are extensions of and results on Signaling Games with infinitely many signals. Already Crawford and Sobel (1982) proved their results for infinitely many signals. Typically, the signals are reals from the unit interval $[0, 1]$. But it is an open question, to the best of my knowledge, whether the results from the general model carry over to linguistic expressions. For language is discrete and we are dealing with countable infinity. Moreover, we'd like to impose special constraints on the signaling games which effect a Montagovian architecture: (i) signals are from a syntactic algebra, (ii) meanings are from a semantic algebra, and (iii) the syntax/semantics-mapping is a homomorphism. Hence, it remains to be seen whether the models of Nowak *et al.* can be extended to recursive rules.

Nevertheless, the new developments create two options to mitigate the *no-structure* objection. The first one tries to extend the existing models by adding recursive rules. But then there will be signals in the game that *should* have a meaning but would never be used (the old meaning-without-use problem; §6.1.5). Can we reapply the lessons learned from Actual Language Relation accounts? Moreover, there are theoretical questions: Are there equilibria? Do they have natural properties? Can, under plausible assumptions, signaling systems evolve? What about the grammars the agents must possess: Must they be the same? If not, how similar must they be? These are open questions one should address in future research. But it

[42] See (Nowak and Krakauer 1999; Nowak et al. 1999a,b; Trapa and Nowak 2000; Nowak et al. 2000; Nowak 2000). Zuidema (2005:§5) also develops an evolutionary model but there are no recursive rules and the set of meanings is assumed to be finite. Recently, also Skyrms (2010) has explored the approach of Nowak *et al.*

seems to be more of a technical problem that will eventually be solved.

The second option does not rely on future research. It reapplies the lessons learned from Actual Language Relation accounts by (i) restricting the explanatory domain to (finite) *effective languages* and by (ii) adding translators as a cognitive component:

(2) Expression e means m in language $\mathcal{L}_{\mathcal{T}}$ of population P iff P uses language $\mathcal{L}$, $\mathcal{L} \subseteq \mathcal{L}_{\mathcal{T}}$, members of P process $\mathcal{L}$-utterances via an $\mathcal{L}$-determining translator $\mathcal{T}$, $\mathcal{T}$ generates $\mathcal{L}_{\mathcal{T}}$, and $\mathcal{T}$ pairs e with m.[43]

According to this proposal, the meanings of the expressions in the effective language $\mathcal{L}$ used by a population P are determined by the conventional (equilibrium) behavior in a (normative) signaling system. $\mathcal{L}$ consists of all the finitely many words and structured expressions that the members of P would possibly use. The meanings of the expressions in the possibly infinite part of $\mathcal{L}_{\mathcal{T}}$ are explained by the translators the members possess. These translators can be realized by a grammar and determine the behavioral dispositions bringing about the conventional signaling regularities.

On the basis of these two options to deal with the *no-structure* objection, I conclude that the *no-structure* objection against Signaling Games is not a knockdown argument against accounts from this paradigm.

7.2.4 Evaluation

As I've argued above, if one is inclined to endorse Millikan's account, one should rather endorse Huttegger's. For it achieves much of what Millikan achieves in a more perspicuous manner, while not strictly following her. For example, it seems that one cannot fully implement Millikan's notion of a proper function in this framework. But to derive signal meanings, we don't need a full implementation of Millikan's notion of a proper function.[44]

Let us now turn to the evaluation of Huttegger's theory. With respect to conventions, the verdict is as in Millikan's case, namely positive.

[43] This is ALR-E′ of §6.1.5.

[44] In an equilibrium of a signaling game, there is a relation between the states and the receiver's reactions. Using this relation, we can "simulate" proper functions: a signal's "simulated" proper function is to indicate the state and/or to demand that the hearer reacts in a certain way. This is sufficient to derive the signals' meanings. The discussion of Lewis' Signaling Games theory in chapter 5 substantiates these claims.

Social norms Huttegger's basic account of social norms is not fully adequate according to my adequacy conditions. Clearly, he had a more modest goal. My interest is to find out how well it does if we use his proposal as an account of social norms. Here are some limitations:

(i) If we understand a normative system as a social norm, the distinction between enforcers, addressees, and arbitrators of a social norm is lost (but required by DesN1). Thereby, the agents that enforce a social norm are the same as its addressees which are, in turn, the same as its arbitrators. We could say that Huttegger assumes an egalitarian social structure in which there are no experts and language rulers.

(ii) Huttegger has to assume that social norms have evolved. While this seems to be true for a class of social norms, this is certainly not true for social norms in general; there are also social norms by decrees.

(iii) Huttegger's theory seems to entail that it is rational or beneficial for people to sanction. Again, it seems to me that this is wrong as a *general* claim (§8.2.4). One reason to think so is that from an evolutionary perspective, sanctioning behavior is elicited by evolved mechanisms. But evolved mechanisms are known to malfunction in environments which they are not adapted to.

One could conclude that Huttegger's basic account of social norms is an account of a special kind of social norm that is interesting since it applies to language use.

Meaning An Evolutionary Signaling Games account à la Huttegger improves on Millikan's account. For (i) the issues relating to the vagueness in the meaning assignment of her account can be addressed and (ii) the problems resulting from endorsing the sentential primacy thesis can be mitigated, as recent developments by Nowak *et al.* show if they are combined with Schiffer's *mental-translator* proposal (DesM6).

But there is no meaning in virtue of social norms (DesM2) since the social norms only have the role of accounting for semantic normativity. Signals get their meanings from the signaling conventions of the underlying game in a normative signaling system.

With respect to semantic normativity, there are still some explanatory gaps. It's left open what it is to demand conformity. Huttegger's theory only entails that it consists in giving a signal in the superimposed game in a normative signaling system which effects a punishment. Such a signal indicates that someone did not conform to the signaling convention and demands to behave conformingly. But how so? To demand something one has to be in a position to do so. Huttegger's theory uses no notion of social structure. So, the best he can say is that everyone is in the position to do so. This still leaves a further question unanswered: How is the nature of the demanding character of these signals explained? Answering this question is also important to explicate a plausible notion of a semantic mistake (DesM3). (Maybe one could use my account of social norms in §8 to do so.)

Finally, as in Millikan's case, Humpty Dumpty's problem remains unsolved (DesM5)

7.3 Outlook for a better theory

In chapter 9, I discuss my alternative conventionalist account. Before I do so, I'd like to (i) bring the big picture back into focus and (ii) to bring out some important properties of meaning for a better theory.

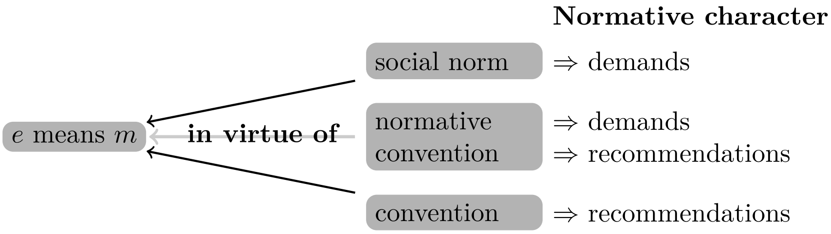
7.3.1 The big picture

In §1.3, I've argued for a conventionalist account that explains meanings in terms of *stable uses*, *i.e.* conventions, social norms, and normative conventions whose pattern of activity consists in the use and understanding of expressions. So far, I've only mentioned normative conventions in passing (§1.2). To remind you, they are a hybrid of a convention and a social norm, and were introduced in §8.2.3. Being a hybrid, they inherit the meaning-determination role of conventions and social norms.

When we turned to the meaning-without-use problem in §6.1.5, I argued for a refinement in terms of *stable linguistic uses* which are stable uses whose behavioral dispositions are realized by a translator (C″), thereby adding a cognitive component.

In chapter 2, I argued that meaning in virtue of social norms is normative (N′). I endorsed the principle that if meaning is normative, then this is because of the way it is determined (MD-ME).

The explanatory architecture that results from these claims can be illustrated as follows:



My claim is that any adequate conventionalist account must adhere to this architecture.[45] A set of problems impose further restrictions on an adequate account.

Solution to Humpy Dumpty's problem Humpty Dumpty's problem requires a solution stating precisely in which ways the meanings of expressions are determined by which members and circumstances (§1.4). This restriction can be addressed by explaining meaning in terms of social norms. For they come with a notion of social structure.

Non-Intellectualism Language use should not only be an Intellectual activity (§6.2.4). This restriction can be satisfied by using an evolutionary notion of convention and by describing language use in a non-Gricean way.

Atomism The problems resulting from the sentential primacy thesis (§6.2.4) and from the existence of language varieties (§6.2.1) suggest that stable linguistic uses should be defined in a more atomistic way relative to language devices (and not just relative to languages).

For this, I propose that for every language device there is a stable linguistic use. According to this proposal, English (whatever it is exactly) consists of a set of stable linguistic uses. Among them is one for "red", another one for "green", and so on.

I take this proposal to be the null hypothesis supported by our folk theory. A rival hypothesis is the *one-convention*-per-language* hypothesis. ("Convention*" is used here in the sense of *stable linguistic use*.) According to the rival, for a person to use English is to be member of *the* convention* to use English. One disadvantage of this rivaling hypothesis is the description of a state of affairs in which people are not fully party to the convention*.

[45] For this reason, it's also immediately obvious why the conventionalist accounts we've considered so far are inadequate: They don't allow for meaning in virtue of social norms.

Typically, language users don't master all of the words of a language. Think of Jonas, an able English user except for the word "apple." He does *not* express his *APPLE* thoughts by using the word "apple" and also does not understand what people do when they utter "apple." Is Jonas a party to the convention* of using English? I take us to be inclined to think so, except for "apple." But what is it to be a party to such a convention* with exceptions? Can one be member *to some degree*? Or be a member *in some respects*?

Here the null hypothesis has more to offer. We can simply describe the case as follows: Jonas is party to the many conventions* of English but not party to the convention* governing the use of "apple". This description also has some explanatory power. Consider a table in which we track the parties to the various conventions* as follows:

Person	"aardvark"	⋯	"apple"	⋯	"zygote"
...	...	...	...	...	...
Jonas	x	x	–	x	x
...	...	...	...	...	...

So, Jonas is a party to the convention* governing the word "aardvark." This can be read off the table by looking at the field value of the row of Jonas and of the column of "aardvark". On the basis of such a description of a possible state of affairs we could postulate the following claim: Person P is competent to use word W if the field of the row of P and of the column W has the value "x"; otherwise not. Such a statement cannot easily be formulated on the basis of the alternative conception.

Moreover, if we endorse Millikan's language independent conception of a linguistic expression, then we can assigns expressions meaning directly and can introduce the notion of a language later in the explanatory order. The case of Jonas above hints at the way it can be done: Languages are clusters of stable linguistic uses prevailing among corresponding clusters of populations. Whether Jonas speaks one of these languages or not can be determined by looking up his entry in the above table. Hence, language varieties can be dealt with easily.

These comparative advantages are what justifies the assumption that for every language device there is a stable linguistic use.

Endorsing the null hypothesis has consequences for stable linguistic uses: Often, they need a certain milieu to flourish in.[46] For one thing, stable lin-

[46] This seems to hold for stable social behaviors in general. In behavioral law and economics,

guistic uses for complex expressions require an environment (or "milieu") in which there are stable linguistic uses for simpler expressions. (An exception is stable linguistic uses for whole utterance types (unstructured signals) that can be used to perform speech acts that do not depend on other conventions or social norms.)

7.3.2 Reflections on meaning

To develop a better conventionalist account, we should ascertain certain general features of an expression's having a meaning in relation to its constitution. The results help us to take informed design choices. To this end, let us return to meaning sentences:

(3) Actually, "appel" presently means *apple* among Dutch users.

I call triples consisting of a group of agents, a time, and a world "coordinates." But why should one think that the truth conditions of simple meaning ascriptions like (4) depend on a coordinate?

(4) "Appel" means *apple*.

The reason for this dependency is simple, as we will see below in more detail: Stable linguistic uses make sentences like (4) true. They exist relative to a coordinate. So, since the truth of sentences like (4) depends on a stable linguistic use existing relative to a coordinate, sentence like (4) depend on coordinates.

But there are subtler features of the conventionalist determination relation. Studying them on the basis of case studies brings out important properties of the relation.

How the case studies work The approach taken here is by performing scenario-based reasoning.[47] To do so, I will typically describe a scenario in

this seems also to be an accepted claim. *Cf.* Posner (2000:4) who points out that legal social norms depend on non-legal background facts (like the agents' acting in a self-interested way and there being other conventions).

[47] I borrow this term "scenario-based reasoning" from Christian Nimtz. As a paradigmatic instance of scenario-based reasoning he takes what philosophers do when they consider Gettier-cases. Nimtz suggests that scenario-based reasoning is a variety of conceptual analysis (Nimtz 2009). Nowadays, it is also used in linguistics, especially in the semantics/pragmatics literature. Scenario-based reasoning is, like conceptual analysis in general,

which an expression is used and understood in certain ways, *e.g.* the word "appel" in Dutch. These scenario descriptions are under a *ceteris paribus*-clause: Unless indicated otherwise, things are in the scenario as they are actually. Then, I will consider the truth of a meaning sentence like (4) or (3). By systematically varying different parts of the scenario the truth conditions of the respective meaning sentences are determined. For example, we could consider the counterfactual situation that only Rotterdammers use the expression to find out how uses of subgroups relate to an expression's meaning something in a "supergroup" (*e.g.* Dutch users).

What I loosely called "parts of the scenario" are in the widest sense things that potentially *fix*, *constitute*, or *determine* an expression's meaning something among the members of a group at a time in a world, where "meaning something" can in turn be understood in two senses, namely in the sense of *something at all* and in the sense of *something in particular.*

The guiding idea is that an expression means something in virtue of there being certain stable linguistic uses. This claim is by now repetitive but it helps us to understand how what can happen to a meaning of an expression relates to these other things. This is to say that we can analyze what can happen to a meaning of an expression in terms of what can happen to stable linguistic uses. Or more technically, if you like, we can analyze the *dynamics* of meaning in terms of *operations* on stable linguistic uses.

The dynamics of meaning consists in four processes: (i) An expression acquires a meaning. (ii) An expression loses a meaning. (iii) An expression maintains its meaning. (iv) An expression changes its meaning. These four processes correspond to the processes operating on stable linguistic use. They also can come into existence, can be maintained, can die out, and can change over time. That this correspondence holds can be explained on the basis of the explanatory architecture of adequate conventionalist accounts: A meaning of an expression in a language is determined by there being a stable linguistic use.

Consequently, (i) an expression e acquires a meaning m if there is a process that results in there being a convention, a normative convention, or a social norm for the expression in virtue of which e means m and e didn't mean m before. Conversely, (ii) an expression e loses a meaning m if e meant

not uncontested. Especially when it comes to far-fetched scenarios, it's not clear that we can reliably reason about them. I try to apply the method with care by focusing on homely cases. What's more, I don't see an alternative way to ascertain the features of meaning I'm interested in.

m before and that in virtue of which e meant m does not exist anymore. Trivially, (iii) e maintains a meaning m if e meant m before and that in virtue of which e meant m is also maintained. Finally, (iv) an expression e changes its meaning from m to m' (with $m \neq m'$) if e meant m before, e means m' now, and that in virtue of which e meant m is that in virtue of which e means m' now. These claims are quite coarse but good enough for our purposes.

Overview I turn now to the following cases which are based on the structure of meaning sentences. I discuss some scenarios relating to social norms and social structure in §9.3.1.

1. **Dependence on groups**: This case relates to both the dependence on groups. The question I try to answer is: "If there is a stable linguistic use of an expression, among whom does the expression mean something?"
2. **Dependence on time and historic meanings**: This case relates to the dependence on time and is about the observation that once an expression has meant something in a community, it's hard for it to lose this meaning. The question I try to answer is: "What is the relation between what an expression historically meant and what it currently means?"
3. **Dependence on worlds**: This case relates to the dependence on worlds and completes the justification that an expression's meaning something depends on coordinates.
4. **Having more than one meaning**: This case is about the observation that expressions can acquire different meanings without losing the others. The question I try to answer is: "How's that possible?"

7.3.2.1 Dependence on groups

If stable linguistic uses determine an expression's meaning, then *among whom* does the expression mean what it does? It seems natural to say that an expression means something *among the members of a group.* It turns out that there is a *difference of scope* between conventions and social norms; social norms can reach out in a way conventions cannot. Let us ask the question: "When do we want to say at all that an expression means something in a group?" Consider the following case:

(5) "lepestele" is now conventionally used and understood in a certain way among the English speaking minority in Amsterdam but nowhere else.

I think we still wouldn't say that "lepestele" means something *in English* or *among English users in general.* But I think that we would want to say that the word means something *among the English speaking minority in Amsterdam.* So, if an expression has a meaning in virtue of a (normative) convention, then it means something among the members of the (normative) convention. But just in virtue of its conventional use and understanding in this group, it doesn't mean something among other agents.

For concreteness, let us consider the group consisting of the English speaking minority in Amsterdam. Among the members of this group, there is no stable linguistic use of "lepestele." For this reason, it would be odd to say that the word means something in this group but there is no stable linguistic use of it among the members of this group. Hence, if there were no stable linguistic use of an expression, then the expression does not have a literal meaning. Thereby, we've justified the claim that an expression's having a certain meaning depends on a group.

Now compare (5) about conventions with the variation (6) about social norms:

(6) There is a social norm to use and understand "heap spraying" in accordance with the meaning *attempt to put a certain sequence of bytes at a predetermined location in the memory of a target process* among English speakers but as before, it's actual use and understanding is so far confined to IT-security specialists.[48]

In this case, the addressees of the social norm are all English speakers. Insofar as the social norm exists, I think we're inclined to say that "heap spraying" means what it does among *all* English speakers. So, in the case of an expression having a meaning *in virtue of a social norm*, it seems that the expression means something among *all addressees* of the social norm, even if not all have (so far) acquired the mental state which is characteristic for the social norm in question, or – now loosely speaking – have acquired the relevant "linguistic knowledge." In extreme cases, a term is only known in this sense by a small minority but still it can be reported to mean something in general. For example, the lexicographer Norman (2002:260) reports about specialist dictionaries in which there is "[p]erhaps a term that is understood by only twenty or thirty researchers worldwide in a particular corner of their field." Yet, it seems that such a term would have a specific meaning in a

[48] See (Wikipedia 09.06.2010) for an explanation of heap spraying.

much bigger group – plausibly in virtue of a social norm.

The foregoing considerations suggest the following generalizations:

(7) Conventions: e means m among the members of G only if members of G are parties to the (normative) convention in virtue of which e means m.

(8) Social norms: e means m among the members of G if there is social norm for e which determines that e means m and is addressed to members of G.

Thus, it can be that e means m among the members of G and G is bigger than the set of those persons who are in the characteristic mental state of the social norm in virtue of which e means m.

Let us reconsider (5) and (6) for another aspect: If there is a stable linguistic use of the word among the members of G, is there also a stable linguistic use of the word among G' which is a subset of G with at least two members? What are the differences between meaning in virtue of a (normative) convention and meaning in virtue of a social norm? It seems to me that the following conditional is a conceptual truth:

(9) If there is a (normative) convention among the members of G and G' is a subset of G with at least two members, then there is also a (normative) convention among the members of G'.

This is so because it's constitutive for a convention that its members have a certain disposition to do their part in its pattern of activity and that it is beneficial for them to do so. If these conditions are satisfied for the members of G, then they are satisfied for subsets of G as well.

Since we can distinguish different groups with respect to a social norm, *e.g.* its enforcers and its addressees, we have to be clear among which subsets a social norm continues to exists. If we shrink the set of enforcers, then it seems that the sphere of influence of the remaining enforcers can become smaller. If we shrink, however, the set of addressees, then it seems the social norm should continue to exist. For if there was a social norm among the original group of addressees, then the enforcers could impose their wills on them. If we took addressees out of this group, then the enforcers would be still in a position to impose their wills since they would remain as powerful as before. Hence it seems plausible to endorse:

(10) If there is a social norm addressed to the members of G and G' is a subset of G with at least two members, then there is also a social norm addressed to the members of G'.

These conditionals have the following consequence:

(11) If e means m among the members of G and G' is a subset of G with at least two members, then e means m among the members of G'.

7.3.2.2 Dependence on time and historic meanings

Does an expression's meaning something depend on time? And how does an expression's having a historic meaning relate to it's having or not having this meaning now? These are the questions to which we turn now.

The answer to the first question, while important, is pretty obvious: An expression's meaning something depends on time. Whether an expression e means m is determined by the stable linguistic uses there are. Stable linguistic uses exist *at a time*. So, an expression e meaning something is determined by the stable linguistic uses there are at a certain time. Moreover, it seems correct to say that if there is no stable linguistic use *at all* of an expression at a certain time, then it doesn't mean anything, in particular not m. Thereby, we've justified the claim that an expression's meaning something depends on time.

But what if an expression was used in the past but is not used anymore? Consider the scenario from §1.3:

(12) The word "camera" was once used in the sense *Apostolic camera* naming the (then existing) treasury department of the papacy. This is not common usage anymore but is still listed as the first entry in Merriam Webster Online Dictionary (2010) and known to some, *e.g.* Hanks (2009:302).

One who is using "camera" in the marginal sense of *Apostolic camera* does not seem to make any linguistic mistake. He can well be called a bragger or a weirdo for it. But this is neither to say that he isn't using a meaningful word nor that he's using it not in accordance with one of its meanings. It seems that "camera" still means *Apostolic camera*, even if there is no convention for it anymore. Yet, it seems that "camera" has not changed its meaning. It seems to be wrong to say: "Then it meant *Apostolic camera* but now it means *device used to take photographs*." Rather, I think, we should say that in the 18th century "camera" didn't have that meaning and that nowadays, the word is ambiguous.

So, "camera" seems to be an expression which was used and understood in the past in a certain way but is nowadays not used and understood anymore in this way. How can it still mean now what it meant back then? A little reflection about the scenario suggests that there is nothing mysterious about it. It seems that a social norm and group knowledge are the keys. The principle seems to be this: If there is a social norm according to which it is allowed to use and understand an expression e now as it was used and understood in the past, then if e meant m in the past, then e means m now in virtue of this social norm. It seems that there is such a *transfer social norm* among English users. For the social norm to be effective, there needs to be group knowledge now that e meant something m in the past.

On this explanation, the meaning *Apostolic camera* of "camera" is determined by a transfer social norm and "selected" by the past use (conventional and/or social norm-governed). The impression that "camera" is not used and understood anymore today was wrong because there are still people who know about its historic meaning and could use the word in accordance with its historic meaning and would understand it if "camera" was used in this way. The explanation is flexible. Even if most language users lack what is required for the use and understanding of an expression, the expression can have an old meaning in a big group. Thereby, we have part of an explanation for the fact that expressions often have many meanings and that it is hard for an expression to lose a meaning.

Moreover, if there weren't any knowledge anymore about a historical meaning of a expression, then it seems that one wouldn't be inclined to say that it still means what it meant. But this is not to claim that it would be wrong to say that. Rather, it would be a claim for which there is no evidence and on that grounds, one should be hesitant to make it. And finally, if we discover (say by finding an old text) that an expression meant something m in the past, then it seems that we would be inclined to say that it meant m then and also to say that it means m now. (Consider the "camera" case: Now you know that it meant *Apostolic camera* back then. You would readily understand it this way if someone were to use "camera" in accordance with this meaning. You might even use it in this meaning in reaction to someone else's use. For this reason, I take us to be inclined to say that "camera" also means *Apostolic camera* now.) These judgments seem to be correct and corroborate the proposed explanation.

Still, it seems weird to claim that any old meaning of an expression is among its meanings today. I'm inclined to uphold my position and explain the weirdness differently: The existence of a transfer social norm does not preclude the existence of a further use-regulating social norm according to which one ought to use and understand expressions in accordance with their current meanings. If there exists such a social norm, the result would be as follows: the expression still has its historic meanings while one ought not to use and understand the expression in accordance with these meanings.

It seems to me that transfer social norms and use-regulating social norms exist among "us". If so, we have a good explanation for the phenomenon of historic meanings. If there is no transfer social norm for an expression, then according to my proposal, the expression doesn't have historic meanings.

The foregoing considerations suggest the following generalizations:

(13) If an expression e meant m in the past, then e means m in virtue of there being a transfer social norm.

(14) A meaningful expression can acquire further meanings without losing or changing its other meanings.

(15) The explanation of an expression having an old meaning in terms of a transfer social norm also explains why it's hard for an expression to lose a meaning.

7.3.2.3 Dependence on worlds

The point here is obvious: stable linguistic uses exist *contingently.* That is, they could have been otherwise. The conventions which exist here could have not existed; this is to say that there is a world in which the conventions don't exist; and likewise for normative conventions and social norms. So, they do not only vary with respect to groups and times but also with respect to worlds. Since an expression's meaning something is constituted by the stable linguistic uses there are, the point also holds for an expression's meaning something. Thereby, we've justified the claim that an expression's meaning something depends on worlds.

7.3.2.4 Having more than one meaning

Maybe the reader wonders why I discuss the topic of having more than one meaning again. The question now is more fundamental: How, in principle, can an expression have more than one meaning?

Let us focus on an expression's meaning something in virtue of conventions. If a word is used and understood in different ways, how can it then still mean anything at all? Shouldn't we expect a conflict in the uses and understandings of the expression which destabilizes the patterns in a way precluding there being a convention? But yet there seem to be many words having many meanings. We've just studied a scenario where a historic meaning was involved. There it seemed that one and the same word acquired more and more meanings ("camera" has also other meanings). Here is a new example:

(16) "fluke" means *the end part of an anchor* and *a stroke of luck*. The word has different etymologies and for this reason, the sound type (or string) "fluke" belongs to different lexemes (Wikipedia 19.02.2010).

So, it seems that if we individuate words not by sound type (or spelling type) but by their lineage, then so far we have had no true example of a word with more than one meaning. For it does not follow that since the *sound type* "camera" has more than one meaning, the *word* "camera" has also more than one meaning. For this reason, it might not be too implausible to assume that most words just have one meaning. The observation is then that at least very often when we say: "This word has more than one meaning", what we mean is that *the so-named sound type has more than one meaning.*

It still seems surprising that one sound type can have more than one meaning. How can there be two stable but different uses and understandings? A plausible explanation is that such uses and understandings are marked in different ways. On the basis of these marks, the two uses and understandings can be told apart. Possible marks are: Different syntactic category (*e.g.* verb and noun), occurrence in different, typically non overlapping topics ("class" as in "school class" or as in mathematics), semantic stereotypicality or compatibility ("blue" as in "he was blue" or "the sky was blue"). So, if two distinct uses and understandings of a sound type are distinct enough so that they can be told apart on particular occasions, they can be stable. If they cannot be told apart, then there is competition. There seem to be many possible outcomes in such a situation: *e.g.* both uses are destabilized and the sound types die out or one of the uses wins at the cost of the other. But at this point we should distinguish two questions: (i) What is the dynamics of such a system? (ii) What determines the meaning of an expression in which way? For the philosophical project, I think only the latter question is relevant. Part of the answer is: However a stable linguistic use of a word came about, it determines the meaning. Arguably, the first question is much harder to answer. The foregoing considerations suggest the following generalizations:

(17) If expressions are individuated by lineages, then it seems that most meaningful expressions just have one meaning.

(18) As long as different uses and understandings of a sound type can be told apart *often enough*, they can be stable and determine different meanings for the type.

(19) We should distinguish the question how a conventional (or social norm-governed) use of an expression came about from the question whether there is a convention or a social norm for an expression. For the conventionalist project, only an answer to the second question is crucial.

7.4 Summary

In this chapter, I digressed into many topics. Let me return to some important points. Millikan's account explains an expression's meaning in terms of the stabilizing function its conventional use has. This plausible idea faces a series of obstacles which can be partly overcome by switching from her account to a theory formulated in the framework of evolutionary game theory. A promising example is Huttegger's theory. Such accounts can be extended to deal with structured signals and hence the *no-structure* objection can be answered. However, such accounts provide so far only basic accounts of social norms. Consequently, semantic normativity cannot be adequately explained.

Hence, a better account is required. To learn from the mistakes of past accounts, I stated a set of problems and ways to solve them. The three key insights are: (i) We need an account of social norms. (ii) We should assign meanings to language devices directly (without referring to a language). That is, we should assume for each language device a stable linguistic use.

Chapter 8

An account of social norms

> [S]ocial norms against noncompliance may [. . .] produce a critical attitude toward rule violators. In such cases, the judgment is not tied to self-interested motivations in a particular social interaction but, rather, to a general criticism of those who violate any social institution.
>
> *Institutions and social conflict*
> Jack Knight

In this chapter I develop an account of social norms. Doing so will close one of the gaps of the conventionalist project. The explicated notion is one that satisfies the pre-theoretic characterization of social norms (§1.2):

S0. A social norm is *social*: there are various not necessarily disjoint groups including (i) a group E of enforcers and (ii) a group G of addressees.

S1. A social norm *involves a pattern of activity*: (i) there is a pattern R of individual activities of the addressees, (ii) R determines for a range of activities whether they are conforming or deviating, and (iii) at least on some occasions, the addressees would behave in a way conforming to R.

S2. A social norm is *prescriptive*: A social norm is, in part, constituted by a norm N which determines for a range of activities whether they are prescribed, forbidden, or allowed. The norm N prescribes to conform to R. N is enforced by the enforcers who accept it and have power over the addressees.

S3. A social norm has a *demanding character*: The enforcers are in a position to demand conformity to R from the addressees.

S4. A social norm is *relatively robust*: in all near futures, the enforcers accept the same norm and at least on some occasions, the addressees behave in a way conforming to R.

On the basis of this characterization, I mostly agree with the quote of Jack Knight (2004:72) used in the epigraph. But I would strengthen the claim: All social norms produce a critical attitude against violators, if we understand "critical attitude" in the sense of a *normative attitude*, as I will propose.

In the remainder of this chapter, I first introduce my notion of a social norm in §8.1. In §8.2 I discuss selected topics with respect to (i) the distinction between recommendations and demands, (ii) the purpose of social norms, (iii) kinds of social norms, and (iv) the relations between conventions and social norms, finally introducing the notion of a normative convention. I evaluate my proposal in §8.3. The chapter ends in §8.4 with a summary.

8.1 A notion of a social norm

To cut a long story short, I define a "rationalistic social norm" as follows (other kinds are defined in §8.2.2):

There is a *rationalistic social norm to conform to the pattern of activity R among the members of a group G (addressees) enforced by members of a group E (enforcers) accepting a system of norms N at time t in world w* iff

RSN1. the enforcers have power over the addressees at t in w;

RSN2. according to N, the addressees are N-required to conform to R and N-forbidden to deviate from R;

RSN3. each enforcer accepts a system of norms of which N is a part at t in w;

RSN4. the enforcers tend to sanction the addressees' R-concerning behavior at least partly because of (RSN3) their accepting N at t in w; and

RSN5. the addressees tend to behave according to R at least partly because of (RSN4) the enforcers' tendency to sanction their R-concerning behavior at t in w.

"R-concerning behavior" in RSN4 is a shorthand for "behaviors that are conforming to R or deviating from R."

The definition is supposed to entail that minimal epistemic conditions are satisfied, namely that the enforcers are in a state of "awareness" of the norms they accept and that the addressees are in a state of "awareness" of the sanctioning behavior. But no stronger epistemic conditions need to be satisfied (*e.g.* mutual/common knowledge/belief). The definition also does not rule out that such stronger epistemic conditions are satisfied in particular cases.

An example of a rationalistic social norm is the rule of the road: The policemen are the enforcers; the drivers are the addressees. The police have (institutional and physical) power over the drivers (RSN1). The policemen accept a system of norms according to which driving right (R) is N-required and deviations from it are N-forbidden (RSN2–3). Policemen tend to fine deviations partly because they accept this system of norms (RSN4). The drivers tend to drive right partly because they could be fined if they deviated (RSN5).

The rule of the road is *also* a convention since there is a good alternative to it (driving left) and since it is beneficial for everyone to drive right on condition the others drive right. Hence, I will say that it is a *normative convention.* The notion is introduced in §8.2.3.

In the next three sections, I elaborate on the clauses of the definition relating to (i) power and social structure (RSN1), (ii) the notion of accepting a norm (RSN2–3), and (iii) behavioral tendencies (RSN4–5).

8.1.1 Power and social structure

According to RSN1, the enforcers have power over the addressees. To make sense of this clause I need to say first what *power* is. To do so, (i) I motivate a popular conception from sociology and social psychology. (ii) In terms of it, I explicate RSN1. (iii) I relate my proposal to a recent theory developed by Peter Gärdenfors (1993) which inspired my proposal.

Power and its sources The notion of power I'm interested in comes from Max Weber and relates to the ability to influence someone else's behavior.[1] In this sense, power is an irreflexive relation between persons. This seems like a plausible start. For it avoids the pitfalls of understanding it as a

[1] Weber defined "power" in different ways. One of his characteristic formulations is: "the ability of a person to impose one's will on someone despite resistance," see Weber's *Theory of social and economic organization* (Weber 1997:152).

property as in "He's powerful" or something very abstract as when we use the noun "power" in "Power is bad."

The Weberian conception of power goes well together with the conception of social structure in terms of a system of social relations. On that understanding, power relations are one of the constituents of a society's social structure. Since social norms entail a power relation, social norms are a constituent of social structure, but not the sole one. For power relations can be constituted merely by physical power. Hence they needn't depend on there being social norms.

In their classic study "The bases of social power", The social psychologists John French and Bertram Raven studied how power relations are constituted and which characteristic differences result from that.[2] Their classification is not uncontested. But it has aged quite well and is widely accepted. For this reason, I rely on their classification:

- **Positional power**: The power an individual has in virtue of being in a certain position within an organization.
- **Referent power**: The power an individual has in virtue of having charisma and interpersonal skill to attract others and build loyalty.
- **Expert power**: The power an individual has in virtue of being skilled or expert about something when the other is in need of these skills or expertise.
- **Reward power**: The power an individual has in virtue of her ability to reward.
- **Coercive power**: The power an individual has in virtue of her ability to punish, also by demoting or by withholding rewards.

Reward power and coercive power create incentives for the agents at which the rewards and the coercion are directed. While it's an interesting question whether people are rather motivated by rewards or punishments, I collapse the two under the name "sanction power" to simplify matters. According to my definition of a social norm, enforcers of a social norm have sanction power.

With regard to positional power, I think there is no reason to consider the organizations to be *formal.* Let us widen the sense so as to include *informal* organizations. Then we can conceive of positional power as one which an individual has in virtue of certain social norms. (I'll use this kind of power to introduce arbitrators in §9.3.)

[2] See (French and Raven 1960).

Moreover, often there are secondary effects. Think of a politician or a medical doctor. In virtue of their institutional status, their power does not extend to ordinary life outside the institutions. Yet, even if they don't have a special status in ordinary life, we often treat them as if they did. Hence social structure can result as a *secondary effect*.

Explication of RSN1 For simplicity, I suggest to understand RSN1 on an individual reading as in RSN1′:

RSN1. The enforcers have power over the addressees at t in w.

RSN1′. For all enforcers e and all addressees g: e has power over g at t in w.

This might be too strong and also too weak. It might be that not all enforcers have power over any addressee. We could weaken the clause by only requiring *most* enforcers to be so or by requiring that they are so on *average*.

RSN1′ is also too weak: What happens if a coalition of addressees challenges the enforcers? If we are to allow for such situations, then it seems that it might well be that a single enforcer is not powerful enough to answer the challenge. So, for more complex situations one might want to consider the following interpretation of RSN1:

RSN1″. The enforcers could form a coalition that would together have power over any coalition the addressees could possibly form.

On the basis of the individual reading RSN1′, it is only required that the enforcers have power over the addressees. In particular, it's not required that the addressees *intend* to conform to the norm the enforcers accept. But it's compatible with RSN1′ that they have such an intention.

One should also not think that the enforcers use vast sanctioning power to make the addressees conform to the social norm's pattern of activity. If the addressees depend on the enforcers in some ways, then it can be that the enforcers never have to use their power.[3] Also, imagined power is power as long it disposes the addressees to act as the enforcers want. But such situations are extremes of what is possible by RSN1′. Everyday situations are likely somewhere in between.

[3] Emerson (1962) provides a theory of the relation between dependency and power.

The effect of a realization of a power relation in a population is that the enforcers may impose their will on the addressees. The definition leaves open how the power relation is realized. It allows for different social power structures. They can be egalitarian, dictatorial, or in-between. In the egalitarian case, the group of enforcers is identical to the group of addressees and (almost) everyone has power over (almost) everyone else. In the dictatorial case, one agent has power over all the others. Crucially, the groups E and G needn't be disjoint and may be identical.

Finally, the fact that in a population a certain power relation is realized constitutes the groups of enforcers and addressees. Since populations can change over time, we should impose some continuity principles so that we can say that it's still the same social norm even if a group has a new member or loses an old one. I'll ignore such complications.

Gärdenfors' social power structures Peter Gärdenfors (1993) provides a mathematical theory to solve Humpty Dumpty's problem on the basis of social choice theory and model theory. He explains what an expression in a public language means in terms of (i) what the members of the relevant population mean by the expression in their idiolects and on (ii) the social power structure in the population. His basic approach is to assume that linguistic meanings in the public language are collections of the individual's idiolectal meanings. Given a population of individuals possessing an idiolect, such a collection is determined by the *social power structures* in the population. Social power structures are subsets of the population which are decisive in the sense that they alone – independent of the rest – suffice to determine the social meaning of an expression. In my terminology, such members are both enforcers and arbitrators.

My account uses Gärdenfors' idea to use power structures to determine group-level outcomes: In clause RSN1 I require that the enforcers have power over the addressees. Thereby, the former can impose their will upon the latter. The will the enforcers have is characterized by the common part of the systems of norms they accept (RSN2–3).

8.1.2 Norm expressivism and accepting a norm

In clauses RSN2 and RSN3 of the definition of a social norm, the norm-expressivist theory of Allan Gibbard (1990, 2003) is used by stating that there is a certain system of norms that the enforcers accept. I'll now introduce Gibbard's theory and then return to these clauses.

Gibbard's theory Three central claims of Gibbard's theory are:[4]

G1. We express normative attitudes with normative statements.

G2. Normative attitudes are a special type of conative mental states.

G3. Since normative attitudes are conative, normative statements do not have a descriptive but an "imperative" character.

A comprehensive presentation of Gibbard's norm expressivism would involve several tasks. I focus on what directly pertains to my application: normative attitudes.

To have a normative attitude is to accept a norm. Usually, norms come in systems. To make this a little bit more substantial, we need to say what a system of norms is and we need to explain what role a state of norm acceptance has in an agent's mental life.

A system N of norms can be characterized by a set of sentences formed by a family of basic predicates "N-forbidden", "N-optional", and "N-required" which apply to alternatives of a certain kind, *e.g.* actions or activities. So, a sentence of a particular system of norms might be that some action a is N-required. More complexe predicates can be constructed from the basic ones. Gibbard considers "N-permitted" defined as "either N-optional or N-required".[5] But what have such abstract systems to do with normative attitudes agents A have? Gibbard's answer is that they are linked by a dyadic relation "A accepts N":

For an agent to accept a system of norms is to be in a special type of mental state. Gibbard suggests that this type is *sui generis* and cannot be reduced to other mental states like beliefs, desires, and intentions. For my purposes, it is not so important whether he is right about the *sui generis* claim. So, I move on. Gibbard characterizes the state of accepting a norm as follows:

> The state of accepting a norm, in short, is identified by its place in a syndrome of tendencies toward action and avowal [...] The syndrome

[4] Following Schulte (2008:183) who provides a systematic presentation of norm expressivism.

[5] See (Gibbard 1990:86 ff.).

> that manifests accepting a norm takes in normative discussion and normative governance. In this normative discussion, in unrestrained contexts, one tends to avow the norm. One tends to be influenced by the avowals of others, and to be responsive to their demands for consistency. Normative governance by the norm is a tendency to conform to it. *Accepting* a norm is whatever psychic state, if any, gives rise to this syndrome of avowal of the norm and governance by it.
>
> (Gibbard 1990:75)

So, the states of accepting a norm have two important characteristics:[6] (i) The states have a role in the causation of actions, namely to act norm-conformingly (Gibbard calls this aspect their "normative governance"). (ii) The states have a role in normative discussions. The first characteristic is a variant of motivational internalism which says that states of norm acceptance have a tendency to bring about norm-conforming behavior. This explains the "normative pressure" of norms, especially, when a norm applies to a circumstance and we are aware of it.

The second characteristic consists of two parts: (i) Such states have the role of *normative avowal* (Gibbard 1990:73) which is to have the disposition to avow the norm in discussions. (ii) Such states have the role of *persuadability* (p. 77) which is to have the disposition to adapt one's own normative attitudes when hearing normative utterances of others in certain circumstances.[7]

This second role explains why groups of humans tend to reach consensus on normative matters. The tendency to reach consensus has also beneficial effects when people have to coordinate (pp. 64–68).

This characterization of a state of norm acceptance is quite vague. Peter Schulte (2009:165 ff.) points out that there is considerable disagreement about the nature of normative attitudes. For definiteness, I'll use Schulte's characterization (which should be understood as an improvement on Gibbard). According to Schulte, normative attitudes are conative mental states which have the following characteristics (indicated by "N-want" below):[8]

[6] Here I am again drawing on (Schulte 2008:141 ff.).

[7] Persuadability is not the same as normative governance. Consider the norm according to which it is forbidden to kill. To act norm-conformingly is to not kill people. But to adopt this norm (by persuadability), is not to act norm-conformingly for the norm does not require or make it optional to adopt it.

[8] See (Schulte 2008:160). Schulte pointed out to me in p.c. that Gibbard has never characterized normative attitudes as explicitly as he did. In particular, the fourth characteristic

NA1. Normative attitudes are logically combinable. Example: Suppose I N-want to clean the kitchen on Sunday. I also N-want to write a letter on Tuesday. Then I N-want to do both.

NA2. Normative attitudes are under consistency pressure. Example Suppose I N-want to clean the kitchen on Sunday and to go to the football match. But only one goal can be realized. Hence I N-want to give up one of my N-wants.

NA3. Normative attitudes are generalizable. Example: Suppose *I* N-want you not to harm other people. Then I N-want *everyone* not to do harm to them.

NA4. Normative attitudes are embedded in a hierarchy of higher-order normative attitudes. Example: Suppose I N-want to clean the kitchen on Sunday. Then I N-want to have this N-want.

Generalizability (NA3) results in normative attitudes of the form: "I N-want to ϕ and I N-want you to ϕ!", for some action ϕ and "you" ranging over the members of some focal group, *e.g.* the addressees of a social norm.[9]

This is of importance to distinguish between moral and non-moral normative attitudes. Both kinds can be constituents of social norms. The moral ones, I suggest, are addressed to *anyone* while non-moral ones are addressed only to a focal group. In case of social norms I propose that the focal group is the group of addressees.

From individual norm acceptance to social norms Let me now connect individual norm acceptances to the definition of a social norm. The idea is that norm acceptance explains why the enforcers tend to sanction the addressees in case of conformity or deviation. This in turn has the role to explain why the addressees tend to conform to the pattern of activity of the social norm in question. This requires that the enforcers accept similar enough systems of norms. I implement this idea by requiring that there is a part they commonly accept:

of normative-attitudes deserves special attention. The characteristic is influenced by Hare (Schulte 2008:137–141, 155). But it seems to me that Schulte's explication is faithful to Gibbard's position.

[9] The form of normative wants is the same as the form of the (ordinary non-normative) wants that Lewis requires for his definition of a proper coordination equilibrium. This might explain why people took his definition of a convention to have a demandingly normative character.

RSN2. According to N, the addressees are N-required to conform to R and N-forbidden to deviate from R.

RSN3. Each enforcer accepts a system of norms of which N is a part at t in w.

RSN2 describes a system of norms N that is the commonly accepted part of the systems of norms the enforcers accept according to RSN3. N is the part according to which one ought to conform to the pattern of activity R and not to deviate from it.

RSN3 requires that (enough) enforcers accept systems of norms which have N as a shared part. There needn't be one system of norms that (almost) all enforcers accept. But whichever system of norms each of them accepts, it must have N as a part.

8.1.3 Behavioral tendencies

The use of Gibbard's norm expressivism gives us a mental characterization of a social norm.[10] But satisfying this condition wouldn't be enough for there to be a social norm. It would allow that there is a social norm which is never conformed to by exhibiting a certain kind of overt behavior. In such a situation, I think it would be wrong to say that there is a social norm. Similar to the case of a convention, I suggest that we also impose behavioral conditions in the definition of a social norm. I do so by including clauses RSN4 and RSN5 in the definition of a rationalistic social norm:

RSN4. The enforcers tend to sanction the addressees' R-concerning behavior at least partly because of (RSN3) their accepting N at t in w.

RSN5. The addressees tend to behave according to R at least partly because of (RSN4) the enforcers' tendency to sanction their R-concerning behavior at t in w.

RSN4 requires that the sanctioning behavior of the enforcers is explained by their norm acceptances. Likewise, RSN5 requires that the addressees' R-conforming behavior is explained by the enforcers' sanctioning behavior. The motivation of these requirements is that the enforcers' and addressees'

[10] Gibbard's notion is not devoid of behavioral consequences. He fixes the functional role of the involved mental states as a partly motivational one. So, accepting a norm in Gibbard's sense normally implies that the agent accepting it exhibits conforming behavior. I use a stronger behavioral condition that relates to the enforcers and addressees of a social norm.

behavioral tendencies should not merely be *contingent* upon the existence of a social norm. The sanctioning behavior should *depend on* the norm acceptances and the addressees' R-conforming behavior should depend on the enforcers' sanctioning behavior.

I explicate *sanctioning behavior* as a kind of behavior that reinforces the addressees to conform to the social norm's pattern of activity. It can be realized in different ways. A few examples are: A teacher might put red marks on the pupil's homework. Someone might correct someone else's behavior. One might praise the behavior of someone.

To achieve some generality, we should not commit ourselves to a specific type of sanctioning behavior. The proposed explication has this virtue. As a further general constraint we should impose the following condition: The same kind of behavior has more or less the same effect among enough addressees. Thereby, a punishment is a punishment for enough addressees. Otherwise, it could be that the sanctioning behavior has to be sensitive to the particular addressee. While it's readily conceivable that such sanctions would be effective, they don't seem to be the kind of behavior which is typically elicited when a social norm is reinforced. Moreover, there is a reason why there are sanctioning behaviors that satisfy the constraint: If the addressees share enough of their psychological makeup, then normally there will be things that are pleasant and unpleasant for all of them.

Degree of conforming behavior With respect to RSN4 and RSN5, there is the question how many of the members of the respective group have to exhibit the respective behavior and how often so. Since we're not guided by a precise theory, my answer is vague. It seems that we should not demand that the relevant agents *always* exhibit the relevant behavior. If we assume that human behavior is by and large guided by economic constraints, then we would expect that the addressees conform to the pattern of activity at least so often that conforming less would yield a lower expected utility (which may in part be determined by the expectation of sanctions). By the same token, we would also expect that the addressees don't conform to the pattern of activity more often if doing so would yield a lower expected utility. Likewise, it seems plausible to assume that the enforcers' frequency to sanction depends on what is at stake. That is, the addressees conform in this sense *often enough.* Likewise the enforcers sanction often enough. But how many of the addressees and of the enforcers do so? On this topic I side with Lewis who proposed “almost all” (Lewis 2002:78 ff.). (Arguably, an “enough of them”-formulation could be defended. But it would lead to complications which are not worth it for my application: (i) The power-relation between the enforcers and addressees would become even more complex. (ii) The interaction with the degree of conformity would make it less clear what the conditions RSN4 and RSN5 amount to.)

8.2 Discussion

According to my definition of a rationalistic social norm, social norms are *socially enforced norms*. This brings several topics into focus: (i) How is the demanding character of social norms explained? (ii) Are there other important kinds of social norms besides the rationalistic social norms? (iii) What are the relations between conventions and social norms? (iv) Which purposes do social norms have?

The two key influences on which my account is based are Allan Gibbard’s theory and the theory of Peter Gärdenfors (1993). For my account of social norms, their theories can be seen as complementary: Gibbard develops a theory of what it is for an individual to accept (a system of) norms. Thereby, the demanding character of social norms can be explained (see below). I extended Gibbard’s theory to what it is for a group to have a social norm by drawing on Gärdenfors’ idea that social power structures determine group-level outcomes. In short: Gibbard explains normativity and Gärdenfors explains the effects of power for the group.

8.2.1 Recommendations and demands

One difference between conventions and social norms is that (i) conforming behavior can be demanded in case of social norm but not necessarily in case of a convention. (ii) In case of a convention, conforming behavior can be recommended but not necessarily in case of a social norm (§1.2).

This distinction is one of the kind of directive speech act that can be performed. The question is how it is to be explained. Linguistically, we can often use the same means, namely uttering an imperative:

(1) You ought to do action A in situation S!

(2) Do A in S!

Since we can use utterances of both kinds to recommend and to demand something, linguistic tests don't seem to be sufficient for an explanation of the distinction. A better explanation is required. The most illuminating analysis of the distinction between recommendations and demands I know is by Peter Schulte (2009), based on his thesis (Schulte 2008).[11]

Schulte (2009:164) proposes that the difference between recommendations and demands can be analyzed using a modification of what he calls "standard speech act theory" that is, the speech act theory by Searle and Vanderveken (1985). Using speech act theory, Schulte argues that the standard analysis does not explain the different *normative* characters of recommendations and demands (pp. 163–164). He agrees with the analysis of demands in terms of a speaker's *wants*, whether they are ordinary wants or normative wants.[12] In contrast to normative wants, ordinary wants needn't satisfy conditions N1–4 of normative wants (§8.1.2) and thus can be inconsistent and non-generalizable.

Schulte extends the analysis for recommendations and demands by suggesting that their difference is (at least in part) due to their different *sincerity conditions*. The sincerity condition of a speech act is the condition that must be satisfied in order for a speech act to count as *sincere*.

According to the proposal of Schulte (2009:163), the sincerity conditions are constituted by the following types of wants:

(3) Recommendations: What I would want to do if I were in exactly the same situation as you.[13]

[11] For alternative analyses of the distinction see chapter 6 (in particular pp. 190–200) in (Schulte 2008). Schulte considers and convincingly rejects alternative analyses in terms of (i) Kant's distinction between categorial and hypothetical imperatives, (ii) Mill's and Gibbard's idea of typical emotional reactions, and (iii) standard speech act theory.

[12] In this context, we should understand "want" in "the widest sense of the term" (Schulte 2009:165), including all sorts of pro-attitudes (or conative states): wants, desires, intentions, and – importantly – normative attitudes.

[13] There are questions as to how one should understand that an advisor "is in exactly the same situation as" her addressee. Gibbard (1990:18 ff.) distinguishes between recommending and advising. The former consists in making best use the limited information one has. The latter is acting in full awareness of every relevant fact. Gibbard suggests that advising is more than rationality demands (and recommendingly normative wants are the wants of rationality). I agree. Hence, to be "in exactly the same situation" as the addressee is then to be in a situation in which one has the same beliefs, desires, *etc.* The advisor can evaluate the addressee's beliefs, desires, *etc.* in terms of their consistency, strength, and feasibility given the addressee's means to realize them.

(4) Demands: What I want you to do.

Schulte calls the wants of recommendations "conditional" and the wants of demands "unconditional." The idea is now that the distinction between the two kinds of wants explains the distinction between recommendations and demands.

The latter distinction is analyzed in terms of the types of mental states sincere directive speech acts express. Following standard speech act theory, the relation is established as follows: A speech act A expresses a mental state M if and only if M constitutes the sincerity condition of A (p. 165). So, recommendations express conditional wants and demands express unconditional wants. Since the wants can be ordinary or normative, the distinction can be made finer: Ordinary recommendations express conditional ordinary wants and ordinary demands express unconditional ordinary wants. Likewise, normative recommendations express conditional normative wants and normative demands express unconditional normative wants. The distinction between the two normative directive speech act types explains the speech act difference between *oughts* with a recommending character and *oughts* with a demanding character.

Applying the analysis to conventions and social norms On the basis of the explanation of the difference between normative recommendations and demands, the difference in the normative character between conventions and social norms can also be explained: In case of a social norm, the enforcers accept a system of norms according to which the addressees are N-required to conform to the pattern of activity R of the social norm. By accepting a system of norms, they have the required normative attitudes to sincerely make normative demands. Moreover, they have power over the addressees. Thus, they are in a position to make normative demands. This needn't be so in case of a convention since the existence of a convention neither implies there being a certain power relation nor that its parties accept such systems of norms.

In case of a convention, however, it's beneficial for each party to a convention to do their part in it. Indeed, it seems plausible to say that it's recommendable to do one's part. At least in case of a rationalistic convention, the parties to it have a want to conform on condition the others do as well. Thus, they have a conditional normative want. Things are slightly different in a case of rationally justifiable and dispositional conventions where it seems plausible to say that its parties needn't have normative wants but only ordinary wants.

In case of a social norm, the normative wants of the enforcers that the addressees conform to the pattern of activity needn't be recommendable.

8.2.2 Kinds of social norms

In case of a convention, it was helpful to distinguish between rationalistic, rationally justifiable, and dispositional conventions (§1.2). An obvious question is whether there are corresponding kinds of social norms, besides the rationalist social norms I've defined.

Rationally justifiable social norms I think there is a coherent notion of a rationally justifiable social norm. Plausibly, such social norms can exist among agents whose psychology is so that they *could* have the required normative attitudes. Moreover, while they're disposed to exhibit patterns which are typical for accepting a social norm, it might be that they don't have them (but could acquire them). If this sounds plausible, we can define a "rationally justifiable social norm" by changing RSN3 of the definition of a rationalistic social norm to JSN3 ("JSN" for "rationally justifiable social norm"):

JSN3. Each enforcer could accept a system of norms of which N is a part at t in w.

Dispositional quasi-social norms I think there is no coherent notion of a *dispositional* social norm. For plausibly, conditions N1–4 that normative attitudes have to satisfy (§8.1.2) imply an rationalistic agent conception. Consider the following example:

(5) "Dumb" agents of a community are disposed to behave in a certain way while some of them are also disposed to sanction conformity and deviations.

In such a scenario, I think we should *not* require that the agents have normative attitudes. For the cognitive mechanisms that explain their dispositions can be simpler than the ones required for normative attitudes.

We can, of course, introduce a related notion which could be called "dispositional quasi-social norms" which results from the definition of a rationalistic social norm by dropping clauses RSN2 and RSN3 and by dropping the "because"-parts of RSN4:

There is a *dispositional quasi-social norm to conform to the pattern of activity R among the members of a group G (addressees) enforced by members of a group E (enforcers) at time t in world w* iff

DSN1. the enforcers have power over the addressees at t in w;

DSN2. the enforcers tend to sanction the addressees' R-concerning behavior at t in w; and

DSN3. the addressees tend to behave according to R at least partly because of (DSN2) the enforcers' tendency to sanction their R-concerning behavior at t in w.

In case of a dispositional quasi-social norm, the observable behavior of the involved agents might be very similar to the observable behavior in case of rationalistic and rationally justifiable social norms. The case can only be settled by examining the agents' attitudes: Are they normative or not? If they aren't, then there can only be a dispositional quasi-social norm among them. If the attitudes are normative, then there can be a rationalistic or rationally justifiable social norm among them.

8.2.3 Relations between social norms and conventions

Separating social norms from conventions creates the impression that the two notions are unrelated. But they are not. In this section, I work through the following sequence of claims: (i) Conventions and social norms can exist independently. (ii) While they can exist independently, the existence of a social norm is compatible with the existence of a convention. I flesh out this idea by introducing the notion of a *normative convention.* (iv) There is a tendency that conventions become normative conventions. (v) The distinction between conventions and social norms helps to clarify the normativity debate about "conventions".

Independence of conventions and social norms Let us define a *counterpart of a social norm (convention)* to be a convention (social norm) which has the same pattern of activity as the original social norm (convention). From the pre-theoretic characterizations of conventions and social norms the following claims follow:

(6) The existence of a social norm does not imply the existence of a counterpart.

(7) The existence of a convention does not imply the existence of a counterpart.

(6) follows since social norms need not have an alternative. (7) follows since the existence of a convention implies neither the existence of a power relation nor that there are people who have the characteristic normative attitudes of social norms.

Normative conventions While conventions and social norms can exist independently, their existence is compatible. To this end, let us introduce the notion of a normative convention which is, in a sense, a combination of a convention and a social norm:

A pattern of activity R is a normative convention among the members of a group G at time t in world w iff

1. R is a conventional pattern among the members of a group G at time t in world w; and
2. there is a systems of norms N such that there is a social norm to conform to R among the members of G enforced by its members accepting a system of norms N at time t in world w.

That is to say, a normative convention is a convention with a social norm that enforces conformity to its pattern of activity with the *special feature* that the enforcers = the addressees (= the arbitrators; §9.3).

Since a normative convention is a convention, (i) there exists a good alternative to the conventional pattern of activity, (ii) it's beneficial to conform for each party to it on condition that the others do, and (iii) it's recommendable to conform (a consequence of (ii)).

Since a normative convention is a social norm, the parties to it can be demanded to conform. Hence, a normative convention has both a recommending and a demanding character.

It seems to me that the *informal* rule of the road[14] and many linguistic rules are such normative conventions.

We could have defined normative conventions differently. For example, instead of the *special feature*, one could have allowed for enforcers that

[14] The informal rule of the road is just an evolved behavior which is governed by a social norm. In contrast, many countries have institutionalized it: The legislators are the arbitrators which are distinct from enforcers (= the citizens having executive power, among them being policemen).

are not addressees. But it seems to me that for the normativity debates concerning (i) conventions and (ii) semantic normativity, this particular notion of a normative convention is the interesting one. (I'll return to the second topic in the next chapter.)

From conventions to normative conventions The conceptual relations between conventions, social norms, and normative conventions are rather weak. On their own, they fail to explain a stronger connection we observe between them: conventions tend to become normative in the sense that (i) agents start to evaluate behavior not only with respect to its conformity but also with respect to its correctness or appropriateness, and (ii), sanction mechanisms begin to be established.[15] In short, conventions tend to become normative conventions.

Establishing the connection is not difficult. We can simply state the following hypothesis:[16]

C-NC. Other things being equal, when a convention becomes entrenched among human members, it becomes a normative convention.

It would be interesting to have an explanation why C-NC is true, if it is. For it has the following consequence for the conventionalist project: As soon as a linguistic convention becomes entrenched, it becomes a normative convention, thereby having a demanding character. Hence, the prediction would be that linguistic conventions have *no* demanding character just so long as they are not entrenched. To me, there seems to be a point to this. Consider someone coining a neologism or a situation where the participants just want to communicate and have established a new means to do so. In such a scenario, it seems that the participants are *thereby* not in a position to make a normative demand; otherwise, it seems that they are.

But it seems that the explanation of the truth of C-NC is not obvious. For example one might think that C-NC could be explained by a tendency of humans to have normative attitudes towards *stable social behavior* (conventions, normative conventions, and social norms):

[15] This has been pointed out with different levels of endorsement *e.g.* by Lewis (2002:97–100) and Ullmann-Margalit (1977:88 ff.).

[16] Huttegger (2007) implicitly takes this hypothesis for granted; §7.2.2.

(8) Other things being equal, humans tend to accept systems of norms according to which the stable social behavior that prevails ought to prevail.[17]

But while (8) would be an interesting explanation, it's implausible. For example, it seems that if there ever has been a Hobbesian state of nature, then people wanted to escape it. And should we really believe that Karl Marx thought that the economic organization of the world he lived in ought to remain as it was? If not, it seems that (8) is wrong. Maybe the thing to do is to look for more moderate explanations, *e.g.* by only trying to explain why it is so in case of linguistic conventions. Be that as it may, I leave C-NC unexplained.

The normativity debate about "conventions" In §1.2 I mentioned a reason for separating conventions from social norms: Accepting the distinction has the advantage of clarifying the normativity debates about "conventions".

I'd like to substantiate this claim now. In my terminology I can say that what is ordinarily called a "convention" is sometimes a convention, sometimes a social norm, and sometimes a normative convention.

Lewis talked about conventions since he thought of them as solutions to recurring coordination problems (chapter 4) which only involve prudential *oughts* which do not have a demandingly-normative character.

Gilbert's examples are mostly about social norms. Two typical example of hers are: (i) "The convention in this department is that we dress formally for department meetings" (Gilbert 2008:6) and (ii) "there's a convention in this community that after a dinner party one sends a 'thank-you' note to one's host" (Gilbert 1983:392 ff.). It seems to me that the best way to make sense of these examples is by saying that they are used to report a social norm.

Finally, what Kemmerling (1976:129 ff.) seems to suggest is that the relevant notion of a convention is the notion of a normative convention. For he wants that what he calls "convention" is such that (i) it has a rational alternative and is such that (ii) each member can demand conformity to it.

[17] Sugden (2004:§8.3) seems to accept a proposal along these lines: If agents have expectations about the others' behavior, then they resent when the others frustrate their expectations by deviating from the prevailing stable social behaviors. Sugden explains such resentments as being a "primitive human response" (p. 154).

This is why I think the proponents in the normativity debate about conventions are talking past each other. Since there are interesting conceptual relations between these notions, I think they are talking past each other in an interesting way.

8.2.4 Purposes of social norms

Conventions are beneficial for their members. In case of social norms, this needn't be so. For the enforcers can impose their will on the addressees if they are powerful enough. Why then are social norms maintained (or, in other words, relatively robust)? To answer the question, we need to say what their purposes are. Their purposes can be classified in two classes, one relating to behavior and the other relating to reasoning.

Social norms as incentives for conforming to patterns of activity Social norms help to bring about or maintain social arrangements, irrespective of how good or bad they are. Such a social arrangement can also consist in something not being the case. In this case, the social norm can be said to help prevent that a certain state of affairs obtains. Since the social arrangement can be desired by one and be disliked by another, social norms can both help to further personal goals as well as help to go against someone's goals.

In particular we can say that a social norm creates a motivation to conform to the pattern of activity of the social norm. For human psychology is such as to avoid punishment and to seek reward.[18] In case of a social norm, the enforcers have normative wants that the addressees behave conformingly. Hence, deviations tend to evoke resentment and conformity praise in their bearers.

In some cases, the punishments and rewards can have a highly regular nature, namely if they are implemented by means of a sanction mechanism. Or they can be flakier or be even suppressed because the sanctioning behavior is outranked by other goals the enforcers have – how much so is an empirical question.

Expectations of punishments and rewards strengthen their effects. Expectation of a punishment tends to evoke evasive behavior. Likewise, ex-

[18] This claim seems to be analytical. Nevertheless, it seems also to be rather well supported by empirical results; see *e.g.* Rolls' *The brain and emotion* (Rolls 2004).

pectation of a reward tends to make behavior which triggers it more likely. Hence, the expectation of the stick and the carrot makes the addressees conform even if no punishment or reward is present in some cases.

Since punishments and rewards change the agents' preferences, we might think of a social norm as a *means* for norm-conforming behavior, or as something that creates an *incentive structure* for norm-conforming behavior. As such we can think of social norms as *transformations of games* (in the game theory sense of game). This view was pioneered by Ullmann-Margalit (1977) and further developed by Bicchieri (2006).[19] Basically, if the rewards and the punishments are high enough, any strategy profile in the untransformed game can become stable.[20]

The general function of social norms as an incentive for conformity to a pattern of activity can be more specialized in certain cases: If the social arrangement of a social norm is a public good, its function is to prevent bad outcomes and/or to yield Pareto efficiency.

The function can also be to shortcut normative reasoning. Sometimes, we can think things through and come to a conclusion after a long deliberation. Social norms can make such conclusions focal and save the cost of thinking something through. For example, if one takes a test and considers cheating, one might just stick to what one ought to do according to the social norm, namely not to cheat. Or one might think it through by considering what kind of person one wants to be, how successful cheating would be, and what the risks are, and reason from there towards the conclusion that one ought not to cheat.

But nothing in the definition of a social norm ties it to a public good. A social norm can simply be an instrument for people with kinky attitudes: a social norm can please perverts who derive joy from exerting power. Or it can still the hunger for submission of masochists. Or it might be instrumental for people desiring to show off by being able to correct others' wrongdoings.

Social norms as a regulation device for coordination Often it is desired to steer a group in a certain direction or to regulate their conduct

[19] See also (Magen 2005) for a recent overview of related research in behavioral law and economics.

[20] A theorem which only depends on punishments has been proved by Boyd and Richerson's *Punishment allows the evolution of cooperation (or anything else) in sizable groups* (Boyd and Richerson 1992).

– in action and attitude. Run-of-the mill coordination games like two person's meeting at one of two places illustrate the importance of regulation in action. But also regulation of attitudes, and of particular judgment, is important. We care about seeing the world in similar ways. We prefer to have people with the same ideology around us and generally dislike people with an ideology which is too different from ours. We care for agreement in judgment. Old and also new cases should be classified in the same ways. (I won't try to explain why we seem to have such attitudes.)

Social norms help us to achieve these ends. They are a regulation device. The nature of normative wants is such as to strive for uniformity among people. In particular, normative wants tend to shape preferences in a certain "matching" way. This is so at the agent-level because normative wants are under a consistency pressure. Moreover, uniformity (and sometimes strong opposition) across communicating groups of agents can be expected because of the avowal of norms (§8.1.2). Their avowal furthers agreement (and sometimes disagreement) in judgment in combination with the consistency pressure. Over time, we can expect that certain ideologies converge while others become more sharply opposed. What wins in the end is hard to predict (but the dynamics of some models is well understood).[21] So, another general function of social norms is to shape reasoning and judgments.

From functions to relative robustness From a speculation on the functions of social norms it's still a long way to an explanation of their relative robustness. One guess would be to think that social norms are somehow beneficial to both the enforcers and the addressees. But this doesn't seem to be true for all social norms. For some of them seemingly don't have a benefit: Why do we pay a tip in a restaurant that we don't expect to visit again and why don't we litter if nobody sees us (Frank 1988)? Moreover, even if having social norms is often beneficial, it does not follow that it's *always* so. In case of an evolved mechanism to have norms, *e.g.* along the lines of Sripada and Stich (2006), we should even expect that such a mechanism "misfires" from time to time by accepting a norm that shouldn't be accepted. For these reasons, I don't want to commit myself to an explanation

[21] See for example the survey of Burke and Young (2009). They discuss evolutionary models of social norms under the label of "local conformity/global diversity effect" (Young 1998): the often observed phenomenon that populations tend to be locally uniform but globally diverse.

of the relative robustness of social norms.

But it follows from the structure of the proposed definition of a social norm that if a social norm is relatively robust, then (i) the power-relation among the enforcers and addresses and (ii) the norm acceptance of the enforcers are relatively robust. Hence, to make social norms relatively robust, we have to strengthen these two conditions accordingly.

8.3 Evaluation

My account of social norms is not fully adequate. For its application to meaning in virtue of social norms and semantic normativity, it is good enough. The account's strength is the explanation of (semantic) normativity. Let's establish the weaknesses by going through the list of desiderata:

DesN1. The account must be faithful to the pre-theoretic characterizations of a social norm (S0–4 in §1.2) and of normativity (N1–3 in §1.1.3).

The account presented in this chapter does not yet fully satisfy DesN1 since arbitrators are ignored. (They will be added in §9.3.) Otherwise the definition of a rationalistic social norm and of a rationally justifiable social norm satisfy the pre-theoretic characterization of a social norm. To explain relative robustness we have to assume additionally that the power relation and the norm acceptances are relatively robust (§8.2.4). These assumptions cannot be explained theory internally and are in need of further justification. Finally, Gibbard's theory explains normativity in an adequate way.

DesN2. The account must provide an answer to the question what (social) norms and normativity are.

Social norms according to my account are complex entities that are individuated both behaviorally (prescribed pattern of activity and sanctioning behavior) and psychically (norm acceptance). Norms are explained on the basis of Gibbard's norm expressivism. They are abstract entities (sets of sentences from deontic logic). Normativity is explained on the basis of Gibbard's theory and Schulte's explanation of the distinction between recommendations and demands.

DesN3. The account must provide a taxonomy of the kinds of (social) norms there are.

In this chapter, I've introduced rationalistic social norms and rationally justifiable social norms. Moreover, I defined dispositional quasi-social norms which are, strictly speaking, not social norms since the enforcers are not required to accept norms. In the next chapter, I'll introduce more kinds: linguistic social norms, expert social norms. This is not a complete taxonomy but covers the important ones for the conventionalist project.

DesN4. The account must provide an answer to the question what kinds of epistemic states are involved in a (social) norm.

According to my account, only minimal epistemic conditions have to be satisfied: enforcers need to be aware of their norm acceptances and addressees need to be aware of the sanctioning behavior.

DesN5. It must be possible that in a human population social norms are created, learned, sustained, and changed.

Plausibly, humans can satisfy the conditions for a rationalistic social norm. Moreover, if a subgroup in a human population starts to have power and its members accept suitably similar systems of norms, they can enforce a certain behavior and a social norm is created. If they lose power and/or they stop accepting these systems of norms, then the social norm ceases to exist.

DesN6. The dynamics of (social) norms must be explained.

I offer no explanation to meet this last desideratum. As pointed out in §8.1.1, I assume that populations remain unchanged. This assumption should eventually be dropped. Since my account is inspired by work in social psychology and social choice theory, it should in principle be possible to provide the desired explanations.

8.4 Summary

In this chapter, I've discussed an account of social norms that conceives of them as *socially enforced norms*. To this end, a notion of a rationalistic social norm was introduced from which other notions were defined, among them normative conventions which share elements of conventions and social norms.

The outcome is that social norms became less similar to conventions. This pointed out that the connections between conventions, social norms, and normative conventions should be studied. But in particular the tendency that conventions become normative conventions resisted an easy explanation.

Chapter 9

An alternative conventionalist account

> Well, this is the mystery package. First, a small anecdote. My sometimes mischievous friend Richard Grandy once said, in connection with some other occasion on which I was talking, that to represent my remarks, it would be necessary to introduce a new form of speech act, or a new operator, which was to be called the operator of quessertion. It is to be read as "It is perhaps possible that someone might assert that ..." [...] Everything I shall suggest here is highly quessertable.
>
> *Meaning revisited*
> Paul Grice

In this chapter, I present an alternative conventionalist account. It uses Millikan's account of conventions and my account of social norms. The latter allows me to illustrate which issues can be addressed on its basis. My conventionalist account does not explain how stable linguistic uses evolve. Rather, the account entails claims that are conditional on there being certain stable linguistic uses. That is, the account explains how meaning is determined by stable linguistic uses, however they came about.

The goal is then to provide an explanation of what makes meaning sentences true which respects the account's main features: it is non-Gricean, word-level, non-Intellectual, and normative.

It is *non-Gricean* in that meaning is not analyzed in terms of Gricean speaker-meaning. It is *word-level* in the sense that for each language device, there is a stable linguistic use which determines its meaning. It is *non-Intellectual* in the sense that linguistic behavior is not explained in terms of rational deliberation. It is is *normative* in the sense that at least some demandingly normative speech acts can be performed with meaning sentences.

The explanation I offer is in terms of meaning-determination claims. The claims are stated as conditionals (tentative version):

(1) If there is a conventional use of description δ indicating content μ in group g, then there is a stable linguistic use of δ indicating μ in g.

(2) If there is a social-norm-governed use of δ indicating μ in g, then there is a stable linguistic use of δ indicating μ in g.

(3) If there is a stable linguistic use of δ indicating μ in g, then δ means μ in g.

For reasons that will become clear in the next section, I ascribe meanings to *descriptions* of expressions and not to the expressions themselves. The first and the second conditional expresses that if there is a certain linguistic convention or social norm for a description, then there is a certain stable linguistic use of it. The third conditional expresses that if there is a certain stable linguistic use of a description, then it has a certain meaning among members of a certain group. (I discuss *normative conventions* only in passing since they behave like conventions with regard to meaning determination and like (linguistic) social norms with regard to (semantic) normativity with the exception that they also have a recommending character.)

The form of these claims indicates another feature of the account: Meaning ascriptions are only relative to a group and not (also) to a language. This is possible since I use Millikan's notion of an expression, that individuates them on the basis of their historical uses (thesis MI1 in §7.1.1). Languages and meaning in a language are introduced later in the explanation.

In another sense, a formal language *is* used in the account, namely to describe the possible language uses of populations. The formal language itself does not explain what the expressions mean among members of the population. The explanatory work is done by the stable linguistic uses of the expressions.

The plan for this chapter is as follows: In §9.1, I discuss general aspects of language and language use. In §9.2, I explain meaning in virtue of conventions. In §9.3, I explain meaning in virtue of social norms. I also offer an explanation of semantic normativity, thereby completing the proposal which I started in chapter 2. In §9.4, the notion of a public language is introduced and applied to Humpty Dumpty's problem; in passing I reapply Schiffer's proposal for the meaning-without-use-problem to my account. The account is evaluated in §9.5. The chapter ends with a summary in §9.6.

9.1 A simple language and its use

To keep the discussion manageable, (i) I introduce a simple language called "ABLE-0." Thereby, I can make precise claims about possible language use and its role in the meaning-determination claims. (ii) I define a notion of derivational complexity. This helps us us to avoid the meaning-without-use problem. (iii) I suggest a general description of language use. Thereby, the communicative patterns of linguistic (normative) conventions and social norms can be described that determine the meanings of expressions (or more precisely: their descriptions). The more technical parts can safely be skipped since the details only matter for the description of the communicative patterns. The details are encapsulated by abbreviations I introduce.

ABLE-0 is a simple language for describing utterance types that is yet A (tiny little) Bit Like English. It consists of structured expressions that have compositional meanings. There are two names "a" and "b", two predicates "I" and "J", negation "–", conjunction "&", and an assertion marker "A:".

The syntax and semantics of ABLE-0 are described by a categorial grammar using the formalism of Herman Hendriks (2001). I assume basic familiarity with categorial grammar and won't introduce the details of the grammar for ABLE-0 but simply provide examples.

The lexicon is defined as in figure 9.1. I treat here conjunction ("&") and negation ("–") as *expressions* and not as functors. The role of the mood marker "A" for assertions is akin to the role of the schematic letter "F" for a name of an illocutionary force in Searle and Vanderveken's formal language for illocutionary acts (Searle and Vanderveken 1985). There, an instance of "F" can be concatenated with a name "p" for a propositional content to represent an illocutionary act, as in "$F(p)$". In ABLE-0, assertion-sentences are syntactically of category a, resulting from a mood marker of category a/s and a sentence of category s.

Some expressions of ABLE-0 are in figure 9.2. The expressions of the language are described by using a syntactic term algebra (*i.e.* a description of an expression is its derivational history). For example expression "Ja" is described as "$F_1(\mathrm{J}, \mathrm{a})$" which expresses that "Ja" is the result of applying a syntactic operation F_1 (forming sentences by concatenating a predicate and a name) to the arguments "J" and "a".[1] (I say more about the semantic

[1] F_2 forms more complex sentences by concatenating expressions of category s/s and a sentence of category s. F_3 forms an expression of category s/s by concatenating a sentence

Word	**Syntactic category**	**Translation**	**Semantic type**
a	n	a'	e
b	n	b'	e
I	s/n	I'	$\langle e, t\rangle$
J	s/n	J'	$\langle e, t\rangle$
–	s/s	$\lambda P.\neg P$	$\langle t, t\rangle$
&	$(s/s)/s$	$\lambda P.\lambda Q.(P \wedge Q)$	$\langle t, \langle t, t\rangle\rangle$
A	a/s	$\lambda P.\lambda x.Bel_x(P)$	$\langle t, \langle e, t\rangle\rangle$

Figure 9.1: The lexicon of ABLE-0

Expression	**Description**	**Semantic value**
Ja	$F_1(\mathrm{J}, \mathrm{a})$	$J'a'$
A:Ja	$F_4(F_1(\mathrm{J}, \mathrm{a}))$	$\lambda x.Bel_x(J'a')$
A:Jb	$F_4(F_1(\mathrm{J}, \mathrm{b}))$	$\lambda x.Bel_x(J'b')$
A:–Ja	$F_4(F_2(–, F_1(\mathrm{J}, \mathrm{a})))$	$\lambda x.Bel_x(\neg J'a')$
A:Ja & Jb	$F_4(F_2(F_3(F_1(\mathrm{J}, \mathrm{a}), \&), F_1(\mathrm{J}, \mathrm{b})))$	$\lambda x.Bel_x(J'a' \wedge J'b')$

Figure 9.2: Some expressions of ABLE-0

values, their representation, and their role below.)

The underlying semantics of the language is extensional (tentative version): The meaning of a name is a thing. The meaning of a one-place predicate is a function from things into truth values. The meaning of a complex expression is compositionally determined by the meanings of its parts and the manner of its composition. In particular, the meanings of negation and conjunction are truth-functional in the usual way.

As the lexicon entries in figure 9.1 show, the semantic value of a name like "a" is represented as a', the semantic value of a predicate like "J" as J'. Corresponding to the syntactic operations $F_1, \ldots, F_4$, there are semantic operations $G_1, \ldots, G_4$, which are all simple lambda-applications.[2] For example G_1 yields a truth value if it is applied to an argument of type e (corresponding to names) and an argument of type $\langle e, t\rangle$ (corresponding to a predicate). Of interest is the mood marker "A" for assertion-sentences. Se-

and an expression of category $(s/s)/s$. F_4 forms mood marked sentences by concatenating a mood marker ("A") and a sentence, adding a colon (":") in between.

[2] Hence we don't need to distinguish between them but it supports the account's illustration.

mantically, it denotes a function from sentences P (representing sentential meanings) into functions from individuals into their believing that P.

The chosen semantics is naive and except for one change I'll leave it at that. The change concerns the meanings: The meanings of complex expressions are assumed to be *structured*. Hence, the semantics is not extensional anymore but *quasi-extensional* since the basic parts of the structure are the extensional meanings of the underlying semantics. I'll use the derivational history of a semantic value to model structured meanings.[3]

Even with this change, the semantics for the language is implausible as a model for a natural language semantics. But the exact nature of the semantics (naive or otherwise) is not crucial for the conventionalist project, the reason being that a conventionalist shouldn't make substantial claims about the right kind of semantics; she uses a semantic theory that is suitable for the pragmatic theory she uses to define what literal meanings are (§1.1.1). In contrast, a conventionalist makes a substantial claim about the determination of an expression's literal meaning. Moreover, the employed categorial grammar framework provides the conceptual resources to define a more plausible intensional semantics.[4] Thereby one could represent also a richer fragment of a natural language in a more plausible way.

Let me now return to the role of semantic values, in particular the ones of assertion-sentences. According to my proposal, semantic values have the role to be *that* which is communicated from speakers to hearers.

In case of successful communication with assertions, I assume that speakers have suitable beliefs and, subsequently, hearers do as well (I return to this below in §9.1.3). That is, first a speaker S believes that P, represented as $Bel_S(P)$, and then a hearer H believes that the speaker believes so, represented as $Bel_H(Bel_S(P))$.

The semantic values of expressions are a common element of these attitudes. In ABLE-0 there are only beliefs, but in general the types of attitudes can vary. To prepare for such a generalization, the semantic values of mood marked sentences need to include the characteristic attitude type, *e.g.* $Bel_x(P)$ for x's belief that P. Moreover, the particular agent having

[3] To keep things simple here, (i) let us model these entities with terms of a semantic term algebra that can be defined in terms of the employed categorial grammar and (ii) let us designate the terms by their result. For example, the term representing the result of applying a semantic composition function G_1 to the values J' and a', which is $J'a'$, is also designated as "$J'a'$" – instead of designating it as "$G_1(J', a')$".

[4] See (Dowty et al. 1992) for a systematic development.

the attitude can vary. So, we need to abstract from them. The abstract semantic value $\lambda x.Bel_x(P)$ (with x ranging over individuals and P over sentential meanings) can play these roles for assertion-sentences. I assume that these values are applied to the respective speaker and hearer in particular situations of communication. On this proposal, the semantic value of an assertion-sentence "A:s", for some sentence s whose semantic value is represented as P, is a function from individuals into their believing that P, *i.e.* $\lambda x.Bel_x(P)$.

From expressions to descriptions Another design choice is that *expressions* aren't directly assigned a meaning. The meanings are assigned to their descriptions. This is so for the usual reasons; otherwise one would run into problems with structural ambiguities and homonyms, as in (4) and (5):

(4) I saw the man in the park <u>with the telescope</u>.

(5) Let's meet at the <u>bank</u>!

Structural ambiguity results from there being different ways to derive an expression. Since descriptions are derivational histories of expressions, this kind of ambiguity is resolved.

To deal with homonyms, lexemes are indexed in the description. With these means we can distinguish the different readings of (4) and (5).

The result of this move to descriptions is that meanings are assigned to them. This is the reason why the meaning-determination claims are formulated in the way sketched in the introduction.

9.1.1 Derivational complexity

Let us define the *derivational complexity* of a description inductively as follows: Descriptions of lexemes have derivational complexity 0 and any syntactic operation which is applied to them adds 1. For example, the complexity of "a" is 0 and the complexity of "$F_1(\mathrm{J}, \mathrm{a})$" is 1.

The tentative proposal here is that a description's derivational complexity *may* be *a* factor in determining its cognitive processing complexity (of both interpretation and production). The proposal is simplistic and naive: Two expressions can have the same derivational complexity but still vary in their cognitive processing complexity. Producing an utterance might be more complex than interpreting it (and *vice versa*) and so on. I only endorse the proposal to illustrate the roles of such a measure in a conventionalist account (speaking sloppily about expressions instead of descriptions):

(i) Introducing a complexity measure allows us to describe the potential use of an expression in a community more precisely and helps to make progress on the meaning-without-use problem: We can use infinite languages $\mathcal{L}$ (like ABLE-0), defined by standard grammars linguists would use to describe the language use of a population. On the basis of a complexity measure, we can define which fragments $\mathcal{L}_f$ can be effective languages of the population. I think there is no reason to assume that a plausible general definition of the notion of an effective language can be provided without using such a complexity measure. The fragments $\mathcal{L}_f$ are always finite. Their respective complement $\mathcal{L}\backslash\mathcal{L}_f$ is the unused fragment of $\mathcal{L}$.

Consider, for example, the syntactic conjunction-rule: it is recursive in the usual way. To delimit the conjunctions in the finite fragment, we restrict the infinite class of conjunctions by imposing the condition that the complexity of conjunctions may not exceed a certain value.

With respect to the meaning-without-use problem, the complexity measure can be used to delineate what can be explained by (possible) uses and understandings of expressions and what has to be explained otherwise. Suppose that a population uses a certain effective language and that the expressions of the complement language are also meaningful. My proposal is that we explain the meaning of an expression in a way that depends on whether the expression belongs to the effective language or its complement. In the first case, it means what it does in virtue of its stable linguistic use. In the second case, its meaning cannot be so explained. I'll return to this topic in §9.4 and propose an explanation in terms of Schiffer's translators.

Moreover, typical conventionalist accounts have the consequence that *any* possible use of an expression would determine its meaning. However, when it comes to very complex and complicated sentences in which a certain expression is used, this seems wrong. For if utterances of an expression are too complex or complicated to be used and understood, then they can't be a determining factor for the meaning of the expression or its parts.

So, one role of a complexity measure is to define the scope of the explanation of meaning in virtue of stable linguistic use by characterizing the class of possible utterances of expressions that can determine their meanings.

(ii) If an expression has a certain complexity, then its stable linguistic use must have a corresponding complexity. Thereby, we make sure that the syntactic structures we assign to expressions are justified. In other words: If the uses of expressions do not exhibit a systematic variety, then we shouldn't assign them more complex syntactic structures than warranted. So, a third role of a complexity measure is to make sure that the meaning-constituting use of an expression satisfies a "sanity"-condition.

9.1.2 SCH-patterns

According to conventionalists, it's the use of expressions that determines their meanings. Hence, a conventionalist needs to describe language use carefully to state the meaning-determination claims in the desired generality and precision. My proposal is that the relevant patterns of activity belong to a certain class which I call "SCH-patterns" (for Speaker-Context-Hearer pattern).

The basic idea is that if expressions having a certain description are used in accordance with their meanings, then they indicate (in the sense of *being a natural sign of*) certain mental contents+. Mental contents+ are used in the subsequent meaning-determination claims as the meanings which are assigned to the descriptions.

I use "(mental) content+" as a technical term whose characterization is stipulated to be as follows: Mental contents+ are concepts, structured concepts (thoughts), and types of attitudes (propositional or not). Hence, mental contents+ are not the same as ordinary mental contents which are typically taken to be the contents of attitudes (and parts of such contents).

Since also types of attitudes are mental contents+, the notion of mental content+ is more inclusive than the notion of mental content; one could say that it is the union of ordinary mental contents and some types of mental states. Thereby, I don't want to propose a new ontology of mental contents but to pursue a purely instrumental goal: to have a single domain of semantic values for ABLE-0, in particular for sub-sentential expressions, sentences, and mood marked sentences (which require types of attitudes).

In this section, I introduce SCH-patterns in relation to particular SCH-events. In the subsequent section, I explain what their role is in my account.

9.1.2.1 SCH-patterns and SCH-events

SCH-patterns are patterns of activity (§4.3.3) which are event types. Event types are functions from particulars into events (in the sense of *event token*).

SCH-patterns are defined by abstracting from the particulars that can vary among a certain class of SCH-events. Thereby, SCH-patterns express what is common to this class of events.

An example of an SCH-event is this: Now in the environment, there are a speaker and a hearer: Suske and Wiske. Suske has the belief that a is J. Suske utters u which is an utterance of the expression "A:Ja" described as $F_4(F_1(\mathrm{J}, \mathrm{a}))$ having a derivational complexity of 2. Wiske hears u and comes to have the belief that Suske believes that a is J.

Three events can be distinguished in an SCH-event:

- The S-event: The event that Suske is now in a certain mental state with a certain content+ which has the belief that a is J as a part
- The C-event: The event that now, Suske is a speaker, Wiske is a hearer, the utterance u is of a certain type that can be characterized by its description $F_4(F_1(\mathrm{J}, \mathrm{a}))$, and the environment is of a certain type
- The H-event: The event that Wiske is now in a certain mental state with a certain content+ which has the belief that Suske believes that a is J as a part

There are several relations between these events. At the very minimum, there is both an inclusion-relation and a temporal ordering: The C-event should include both the S-event and the H-event and the H-event should be temporally after the S-event.

Many such events can take place. We can describe them in a suitable regimented language having predicates such as "x is a speaker", "x is in mental state μ_x", *etc.* Thereby, we make explicit what can vary among SCH-events: the speaker S, the hearer H, the utterance u (characterized by its description), the content+ μ_S of the (total) mental state the speaker is in, the content+ μ_H of the (total) mental state the hearer is in, a vector $\vec{t}$ of the time intervals t_S, t_C, and t_H of the respective S-, C-, and H-events, a set V of states of affairs about the environment, and a world w in which the events takes place. Consequently, we can represent SCH-events by tuples of the form $\langle S, H, u, \mu_S, \mu_H, \vec{t}, V, w \rangle$.[5]

[5] SCH-events can be abnormal. There might be no speaker (*e.g.* in case of a written utterance from an unknown writer). There might be no hearer. There might be no utterance (but the hearer believes erroneously that the speaker said something). Such events are represented by setting the respective element in the SCH-event tuple to $\emptyset$.

The example above can be represented as follows:

$$\big\langle \text{Suske}, \text{Wiske}, F_4(F_1(\text{J}, \text{a})), [\cdots Bel_{\text{Suske}}(J'a') \cdots], \\ [\cdots Bel_{\text{Wiske}}(Bel_{\text{Suske}}(J'a')) \cdots], (t_S, t_C, t_H), V, w\big\rangle$$

where "$[\cdots Bel_{\text{Suske}}(J'a') \cdots]$" is the description of the content+ of the total mental state of Suske; likewise for Wiske's content+ which is a belief about what Suske believes. The set V of states of affairs is the set of states of affairs consisting in there being this-and-that stable linguistic use.

By abstracting from these variables and imposing constraints on them, SCH-patterns can be defined.

For ABLE-0 we need SCH-patterns of two kinds: (i) for descriptions δ of expressions and (ii) for families of descriptions of expressions, defined by a syntactic function F_i from which descriptions can be derived by applying it to possible arguments of it (the arguments and the resulting descriptions are always restricted to having a maximal derivational complexity).

I'll provide definitional schemata for both kinds of SCH-patterns below. But for now suppose that we have defined the SCH-patterns for ABLE-0. Then we can stipulate in terms of them what expressions of the following form mean:

(6) the SCH-pattern for description δ indicating content+ μ with derivational complexity of z in environment V before time t in group g in world w

(7) the SCH-patterns for (syntactic) function F_i indicating a matching (semantic) function G_j with derivational complexity z in environments V and V' before time t in group g in world w

Let us ignore the details for the moment and focus on the roles of instances of (6) and (7) (I'll return to these points again below): Instances of the form of (6) and (7) designate particular SCH-patterns. They are used to describe stable linguistic uses which determine the meaning for a description δ of an expression and for a syntactic function F_i for (descriptions of) complex expressions, respectively.

SCH-patterns for descriptions are used to describe the stable linguistic uses of expressions having the description δ whose utterances indicate the mental content+ μ and whose utterance types have a derivational complexity not exceeding z.

Example: A part of such a pattern applying to the expression "A:Ja" is shown in the first line of figure 9.3. The SCH-pattern's description δ is

S	**C**	**H**
$Bel_S(J'a')$	$F_4(F_1(\mathrm{J}, \mathrm{a}))$	$Bel_H(Bel_S(J'a'))$
$Bel_S(\neg J'a')$	$F_4(F_2(\text{-}, F_1(\mathrm{J}, \mathrm{a})))$	$Bel_H(Bel_S(\neg J'a'))$
$Bel_S(J'a' \wedge J'b')$	$F_4(F_2(F_3(F_1(\mathrm{J}, \mathrm{a}), \&), F_1(\mathrm{J}, \mathrm{b})))$	$Bel_H(Bel_S(J'a' \wedge J'b'))$

Figure 9.3: Parts of SCH-events (or -patterns)

$F_4(F_1(\mathrm{J}, \mathrm{a}))$. The indicated content+ μ is $\lambda x.Bel_x(J'a')$. The derivational complexity z is 2. The derivational complexity is used to define the stable language uses that must be part of the environment V for the pattern itself to be a stable linguistic use. Thereby it is required that there is a stable linguistic use of descriptions of mood marked sentences (i) of which δ is a part and (ii) whose derivational complexity does not exceed 2, *e.g.* that there is a stable linguistic use of "A:Ja".

SCH-patterns for functions (of the kind of (7)) are used to describe the stable linguistic uses of complex expressions whose derivational complexity does not exceed z, used in utterances of expressions whose derivational complexity is also not exceeding z, that result from applying the syntactic function F_i to possible arguments, and whose indicated content+ is determined by applying the semantic function G_j to the semantic values of the arguments. The role of the set V is as above: it is used to require that certain stable linguistic uses exist in the environment. The role of the set V' is to ensure that there is a variety of stable linguistic uses of complex expressions resulting from applying F_i to possible arguments of it.

Example: An SCH-pattern for function F_4 indicating G_4 with derivational complexity 3 applies to expression like: "A:Ja" and "A:–Ja".

9.1.2.2 SCH-patterns for descriptions

A description of an SCH-pattern consists of five parts: an abstraction $\underline{\text{A}}$, a description $\underline{\text{C}}$ of the context-event C, a description $\underline{\text{S}}$ of the speaker-event S, a description $\underline{\text{H}}$ of the hearer-event H, and a description $\underline{\text{R}}$ of the relations between the events.

The *SCH-pattern for a description δ indicating content+ μ with a derivational complexity z in an environment $V_{a:\delta,z}$ before time t in group g in world w* is the event type that results from an SCH-event $\langle S, H, u, \mu_S, \mu_H, \vec{t}, V, w\rangle$ by

$\underline{\text{A}}$ (i) *abstracting* from the speaker S and the hearer H, the utterance u, the contents+ μ_S and μ_H of their mental states, the times $\vec{t}$ of the S-, C-, and H-events (all before t), the set V of states of affairs about the environment, the world w in which the events take place; and (ii) *quantifying* over all positions i at which δ can occur in a description:

$\underline{\text{C}}$ such that the description of the C-event is satisfied: At time t_C in world w: (i) $S \in g$ is a speaker and $H \in g$ is a hearer, (ii) the description of the uttered expression u contains δ as a part at position i, (iii) the derivational complexity of the description of u does not exceed z, (iv) $V_{a:\delta,z}$ is a subset of V and at least some state of affairs in $V_{a:\delta,z}$ obtained, and (v) S utters u;

$\underline{\text{S}}$ such that the description of the S-event is satisfied: At time t_S in world w: (i) S is in a mental state having μ_S as its content+, (ii) μ is a part of a belief of S that is a part of μ_S, (iii) the part δ at position i of the description of u indicates this part μ of the content+ μ_S of S's mental state;

$\underline{\text{H}}$ such that the description of the H-event is satisfied: At time t_H in world w: (i) H is in a mental state having μ_H as its content+, (ii) μ is a part of a belief of H about a belief of S that is a part of μ_H, (iii) the part δ at position i of the description of u indicates this part μ of the content+ μ_H of H's mental state; and

$\underline{\text{R}}$ such that the description $\underline{\text{R}}$ expressing the relations between the events is satisfied: (i) the event described in $\underline{\text{C}}$ includes the event described in $\underline{\text{S}}$, (ii) the event described in $\underline{\text{C}}$ includes the event described in $\underline{\text{H}}$, (iii) the event described in $\underline{\text{H}}$ is temporally after the event described in $\underline{\text{S}}$, and (iv) times t_C, t_S, and t_H are before (or at) time t.

In the description $\underline{\text{C}}$, a set $V_{a:\delta,z}$ is used. It consists of states of affairs about the environment. It's role is to secure that if there is a stable linguistic use of something that can be part of a mood marked sentence (defined by the syntactic/semantic functions F_4/G_4), then there are at least some mood marked sentences such that (i) δ is a part, (ii) their description has a derivational complexity of maximally z, and (iii) a stable linguistic use exists for their description. Hence, the role of $V_{a:\delta,z}$ is to rule out that there is a stable linguistic use of a description of a part-utterance type without there being stable linguistic uses for the descriptions of whole utterance types containing the word's description. Thereby, we make sure that the stable linguistic uses for lexemes exist in a suitable environment.

9.1.2.3 SCH-patterns for functions

An SCH-pattern for a complex expression can be characterized by (i) a syntactic function F_i, (ii) a number z, (iii) a matching semantic function G_j

determining the meaning of the expressions resulting from applying F_i, (iv) a set $V_{a:F_i,G_j,z}$ of states of affairs about the environment, (v) another set $V_{i(F_i,G_j,z)}$ of states of affairs about the environment, (vi) a time t before (or at) the instantiations of the SCH-pattern take place, (vii) a group g from which the speakers and hearers are drawn, and (viii) a world w in which the instantiations of the SCH-pattern take place.

The number z expresses both the maximal derivational complexity of the descriptions which may result from applying F_i and the derivational complexity of the descriptions of whole utterance types in which the resulting descriptions can occur as parts.

A semantic function G_j *matches* a syntactic function F_i if their arguments obey the homomorphic mapping between the system of syntactic categories and semantic types (see (Hendriks 2001) for details).

Again, the patterns have existential requirements about the environments. $V_{a:F_i,G_j,z}$ has the role of securing that at least for some description δ indicating a content+ μ that is obtained by F_i and G_j , there is a stable linguistic use of a description δ' of a mood marked sentence (whose derivational complexity does not exceed z) such that δ is a part of δ'.

In addition, there is a second set $V_{i(F_i,G_j,z)}$. It secures that there are stable linguistic uses for all the possible arguments δ_k of F_i which indicate a respective content+ μ_k. This is required to determine the contents+ of the complex descriptions that result from applying F_i to the arguments. A content+ of a complex description is then what results from applying G_j to the respective contents+.

The definitional scheme of SCH-patterns for functions is more complex than the one for SCH-patterns for descriptions. In the instances of the scheme, it is quantified over descriptions and contents+ of possible arguments of the respective functions.

Thereby, the SCH-pattern for a function is more inclusive than the SCH-pattern for a description. The class of SCH-events of a certain SCH-pattern for a function F_i consists of SCH-events whose utterance satisfy *any* of the possible descriptions resulting from applying F_i to possible arguments.

The *SCH-pattern for a (syntactic) function F_i indicating a matching (semantic) function G_j with a derivational complexity z in environments $V_{a:F_i,G_j,z}$ and $V_{i(F_i,G_j,z)}$ before time t in group g in world w* is the event type that results from an SCH-event $\langle S, H, u, \mu_S, \mu_H, \vec{t}, V, w\rangle$ by

$\underline{\text{A}}$ (i) *abstracting* from the speaker S and the hearer H, the utterance u, the contents+ μ_S and μ_H of their mental states, the times $\vec{t}$ of the S-, C-, and H-events (all before t), the set V of states of affairs about the environment, the world w in which the events take place; and (ii) *quantifying* over possible arguments $\delta_0, \ldots, \delta_n$ of F_i and matching contents+ $\mu_0, \ldots, \mu_m$ of G_j (yielding descriptions $F_i(\delta_0, \ldots, \delta_n)$ (= δ) with contents+ $G_j(\mu_0, \ldots, \mu_m)$ (= μ), and quantifying over all positions i at which δ can occur in a description:

$\underline{\text{C}}$ such that the description of the C-event is satisfied: At time t_C in world w: (i) $S \in g$ is a speaker and $H \in g$ is a hearer, (ii) the description of the uttered expression u contains δ as a part at position i, (iii) the derivational complexity of the description of u does not exceed z, (iv) $V_{a:F_i,G_j,z}$ is a subset of V and at least some state of affairs in $V_{a:F_i,G_j,z}$ obtained, (v) $V_{i(F_i,G_j,z)}$ is a subset of V and the state of affairs in $V_{i(F_i,G_j,z)}$ obtained (such that there are stable linguistic uses for descriptions $\delta_0, \ldots, \delta_n$ indicating contents+ $\mu_0, \ldots, \mu_m$), and (vi) S utters u;

$\underline{\text{S}}$ such that the description of the S-event is satisfied: At time t_S in world w: (i) S is in a mental state having μ_S as its content+, (ii) μ is a part of a belief of S that is a part of μ_S, (iii) the part δ at position i of the description of u indicates this part μ of the content+ μ_S of S's mental state;

$\underline{\text{H}}$ such that the description of the H-event is satisfied: At time t_H in world w: (i) H is in a mental state having μ_H as its content+, (ii) μ is a part of a belief of H about a belief of S that is a part of μ_H, (iii) the part δ at position i of the description of u indicates this part μ of the content+ μ_H of H's mental state; and

$\underline{\text{R}}$ such that the description $\underline{\text{R}}$ expressing the relations between the events is satisfied: (i) the event described in $\underline{\text{C}}$ includes the event described in $\underline{\text{S}}$, (ii) the event described in $\underline{\text{C}}$ includes the event described in $\underline{\text{H}}$, (iii) the event described in $\underline{\text{H}}$ is temporally after the event described in $\underline{\text{S}}$, and (iv) t_C, t_S, and t_H are before (or at) t.

Two comments are in order: (i) To define the descriptions generated by the (syntactic) function F_i, it is quantified over possible arguments $\delta_0, \ldots, \delta_n$ of F_i and over possible arguments $\mu_0, \ldots, \mu_m$ of G_j in the $\underline{\text{A}}$-part. (ii) A so-defined description δ generated by F_i has the form "$F_i(\delta_0, \ldots, \delta_n)$" with contents+ of the form "$G_j(\mu_0, \ldots, \mu_m)$" and is used in the $\underline{\text{C}}$-, $\underline{\text{S}}$-, and $\underline{\text{H}}$-part of the scheme. Otherwise, the form of the scheme for SCH-patterns for a function is the same as the one for SCH-patterns for descriptions.

9.1.2.4 The type-token relation

SCH-events are tokens of an SCH-pattern if the description of the utterance in the C-event of the SCH-event is matched by the description of the utterance type of the SCH-pattern; other things needn't match. If all the things match, then we say that the SCH-event *conforms* to the SCH-pattern.

An SCH-pattern describes the roles of both speakers and hearers in SCH-events that are tokens of the pattern. Thereby, we can define for both speakers and hearers in an SCH-event what it is to conform and what it is to deviate from the pattern. For a speaker, to conform to an SCH-pattern is to only utter a certain expression if one is in a mental state whose content+ has a part that is indicated by the utterance. For a hearer, to conform to an SCH-pattern is to come to believe, upon hearing the utterance, that the speaker has a certain attitude which has a certain part that is indicated by the utterance. Deviations are defined in terms of non-conformity.

9.1.3 Language use in terms of SCH-patterns

Plausibly, normal linguistic transactions can be described as SCH-events. One could express what is going on in such linguistic transactions by saying that the speaker transfers a mental content+ to the hearer's "belief box" by producing an utterance.

A content+ that is transferred from a speaker to a hearer is a part of the content+ of the speaker's total mental state and a part of the content+ of the hearer's total mental state.

For the SCH-patterns I consider, I assume that in a case of successful linguistic communication, the transferred content+ of the speaker is identical to the content+ in the hearer's belief box.[6]

According to this proposal, in case of successful communication a hearer learns something about the state of the speaker. To learn something about the world, the hearer has to draw the further "unboxing"-inference of the kind "*Ceteris paribus*, if I believe that the speaker believes that P, then I believe that P." Among the *ceteris paribus* conditions are the speaker's sincerity, her reliability with regard to P (or in general), her reputation, and different sorts of circumstantial conditions (*e.g.* conditions about the speaker's belief acquisition).

The talk of identity of a part of the content+ requires us to conceive of contents+ as something structured. I think we should conceive of them as structured propositions with an attitude marker (where such a marker is

[6] One could weaken the identity assumption by assuming that the corresponding contents+ are similar enough.

something that is a characteristic feature of attitudes of the respective type in the mental life of agents of a certain kind). So, both the contents+ as well as the utterances are typically structured.

The parts of the SCH-events in figure 9.3 illustrate this: A speaker S is in a mental state whose complex content+ (has a part that) is a propositional attitude – represented in the S-part of the pattern. S's complex utterance indicates the attitude – represented in the C-part. Upon hearing the utterance, a hearer H comes to be in a mental state whose complex content+ (has a part that) is a propositional attitude – represented in the H-part.

Derived SCH-patterns and derived stable linguistic uses SCH-patterns for functions have a generality which SCH-patterns for descriptions do not. For the former are defined for a variety of descriptions while each of the latter is defined only for a single description. We can derive SCH-patterns for descriptions from SCH-patterns for functions. For example an SCH-pattern for the syntactic conjunction-rule defines a variety of descriptions that differ in the number of conjunction occurrences. The SCH-patterns for descriptions that can be derived from the SCH-pattern for the syntactic conjunction-rule are the ones whose description contains a certain fixed number of conjunction occurrences.

This has consequences for the use of SCH-patterns as the patterns of stable linguistic uses (*i.e.* a certain kind of (normative) conventions and social norms; §6.1.5): If in a group a stable linguistic use exists which has a certain pattern of activity, then this pattern must have been realized in this group. This is not to say that every possible realization of the pattern must have been realized but at least in some situations some possible realizations must have been realized. Now consider the following claim:

(8) If in a group g a stable linguistic use exists whose pattern is a certain SCH-pattern Θ for a function F_i, then for all SCH-patterns Θ_d for descriptions that can be derived from Θ: in g, a stable linguistic use exists which has Θ_d as its pattern.

Clearly, (8) is false. Derived SCH-patterns for descriptions needn't have been realized. So, we need to distinguish between stable linguistic uses *simpliciter* and *derived* stable linguistic uses. For the latter, we don't assume that their patterns have been realized. On the basis of this distinction, I make the following assumption:

(9) If in a group g a stable linguistic use exists whose pattern is a certain SCH-pattern Θ for a function F_i, then for all SCH-patterns Θ_d for descriptions that can be derived from Θ: in g, a derived stable linguistic use exists which has Θ_d as its pattern.

Subsequently, I use "stable linguistic use" in the inclusive sense of a stable linguistic use *simpliciter* or a derived one.

The role of SCH-patterns in the meaning-determination claim Below, I'll state the meaning-determination claims for meaning in virtue of conventions and meaning in virtue of social norms, in terms of stable linguistic uses whose patterns of activity are SCH-patterns. Thereby, SCH-patterns play an important role in these claims. They determine what conforming and deviant uses and understandings of expressions are. Social norms and (normative) conventions add (i) a group in which these uses and understandings prevail, (ii) their stability, and (iii) their arbitrariness.[7]

Since stable linguistic uses are relatively robust, deviant SCH-events in a group can be tolerated up to a degree that depends on the particular (normative) convention or social norm. In case of conventions, the degree depends on the beneficiality of the conforming behavior. In case of social norms, it depends on the will and power of the enforcers.

So, deviations from SCH-patterns that determine a description's meaning shouldn't bother us too much as long their corresponding stable linguistic uses continue to exist. If they don't, then according to the central meaning-determination claims, the respective descriptions don't have a meaning.(They can still have a meaning by means of a transfer social norm; see §7.3.2.2).

More problematic is the following objection: Members of a community of language users always believe that they will enjoy a splendid afterlife (P) and they believe tautologies (Q). But if so, any utterance of a mood marked sentence a among the members of the community also seems to indicate P and Q. But it would be an implausible result of my account if it entails in such scenarios that mood marked sentences not only mean what they do but also P and Q.

[7] While social norms in general needn't have an alternative, linguistic social norms can be assumed to have alternatives since there could have been other form-meaning pairings.

Reply: We have to distinguish in virtue of what a means what it does.

Case “meaning in virtue of conventions”: The convention that has an SCH-pattern for the description of a has a certain proper function (§7.1). Having this proper function explains why the convention prevails. Hence, my account entails that if to indicate P and/or Q is part of this proper function, then description of a means also P and/or Q. This seems plausible. But a stronger claim can be justified: Arguably, to indicate P or Q is not part of this proper function. For indicating P or Q is not comparably beneficial for the members of the community since they have the respective beliefs, whether they utter mood marked sentences or not.

Case “meaning in virtue of social norms”: The social norm that has an SCH-pattern Θ for the description of a is enforced by enforcers accepting systems of norms which have a system N of norms as a part (see below in §9.3). According to N, it is N-required to conform to Θ and N-forbidden to deviate from Θ. Hence, it depends on whether N is such that conformity to Θ amounts also to indicate P and/or Q or not. If so, then a mood marked sentence s also means P and/or Q among the addressees of the social norm; otherwise not. This seems to me to be a plausible claim.

SCH-patterns are non-Gricean (and also different from Millikan’s)
SCH-patterns describe a certain co-occurrence between (i) events in which the speaker utters something, (ii) events in which the speaker is in a mental state whose content+ has (as a part) a certain content+, and (iii) corresponding events in which the hearer comes to be in a certain mental state whose content+ has, as a part of one of her beliefs about the speaker, the same content+ as the respective part of the speaker’s content+. The relation between the utterances and the contents+ is one of indication. Neither speaker intentions nor the hearer’s recognition of them are used in the description of SCH-patterns. Hence, SCH-patterns are not described in Gricean terms. Since also the stable linguistic uses in terms of these patterns are non-Gricean, the account is non-Gricean.

Indication is usually considered to be too weak for Gricean analyses of speaker-meaning. This is unproblematic for my account, as far as I can tell. For speaker-meaning is not used in the meaning-determination claims.

The SCH-patterns are similar to the ones of Millikan’s account (§7.1), if suitably restated, with the exception that in case of assertions, the hearer reaction is not to come to believe that P (for some P) but to come to believe that the speaker believes that P. I've made this change to avoid problems we would run into when we extended the account to other speech act types. For example, if someone promises to do something, then plausibly, the speaker intends to do so. But we shouldn't say that in this case, the linguistic understanding of the hearer consists in having or coming to have the intention to do so.[8] More plausibly, in case of promising the linguistic understanding consists in believing that the speaker intends that P (or coming to believe so). So, the advantage of stating the hearer-part of SCH-patterns in this way is that it generalizes to other speech act types:

Let Ψ denote a certain type of attitude (like belief, desire, knowledge, or intention) and let M_Ψ be a mood marker in the language indicating attitudes of type Ψ. Then we can say in analogy to assertion-sentences of the form "A:s" that "M_Ψ:s" is a sentence mood marked with M_Ψ. Such a sentence is aptly called a M_Ψ*-sentence*.[9] The semantic values of M_Ψ-sentences are represented as $\lambda x.\Box_x^\Psi(P)$ where x ranges over individuals, $\Box^\Psi$ represents Ψ (*e.g.* Bel, Des, K, Int), and P is a sentential semantic value.

Hence, if speaker S indicates attitude Ψ with content P by uttering a M_Ψ-sentence $M_\Psi : s$, then the semantic value of $M_\Psi : s$ is $\lambda x.\Box_x^\Psi(P)$ and the successful linguistic understanding of $M_\Psi : s$ by hearer H consists in $Bel_H(\Box_S^\Psi(P))$.

9.2 Meaning in virtue of conventions

In this section, meaning in virtue of conventions is explained. Consequently, I have to elaborate on what I take linguistic conventions to be and how they determine an expression's meaning. I use Millikan's account of conventions (chapter 7).[10]

The basic idea of the meaning-determination claim is as follows: If speakers and hearers conventionally use and understand an expression, then utterances of the expression indicate a certain mental content+. In virtue of the conventional use and understanding, this mental content+ is the expression's meaning among the members of the convention.

We can cash out this idea in terms of SCH-patterns. Suppose an SCH-

[8] I thank Lars Dänzer for making me aware of this issue which was present in a prior version of my account.

[9] The notation is borrowed from (Grice 1969:171 ff).

[10] With minor modifications, one could also use a different account of conventions.

pattern for a certain description δ indicating a certain content+ μ with a certain derivational complexity is a convention among members of some group (the other elements, such as the environment, are left out for readability). Then the claim is that among the members, the description δ had the meaning μ.[11] If so, linguistic conventions are conventions whose pattern of activity is an SCH-pattern. Hence, I put forward the following thesis:

C-M. If there is a conventional SCH-pattern for description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among members of g at t in w.

To illustrate, suppose the description "$F_1(\mathrm{J}, \mathrm{a})$" actually means $J'a'$ among members of some group g. On the basis of my account, this could be so if there were a conventional SCH-pattern for description "$F_1(\mathrm{J}, \mathrm{a})$" indicating $J'a'$ with derivational complexity of at least 2 among members of g now in this world (the complexity must be at least 2 since there must be a use of mood marked sentences whose complexity is at least 2; see condition (v)). Since the description "$F_1(\mathrm{J}, \mathrm{a})$" is complex, there can only be a conventional SCH-pattern for it, if among members of g now in this world: (i) there is a stable linguistic use whose SCH-pattern for F_1 indicates some semantic composition function G_1, (ii) there is a stable linguistic use whose SCH-pattern for "J" indicates J', (iii) there is a stable linguistic use whose SCH-pattern for "a" indicates a', (iv) applying F_1 to "J" and "a" yields "$F_1(\mathrm{J}, \mathrm{a})$" and applying G_1 to J' and a' yields $J'a'$, and (v) there are some stable linguistic uses whose pattern applies to mood marked sentences whose description has $F_1(\mathrm{J}, \mathrm{a})$ as a part, *e.g.* $F_4(F_1(\mathrm{J}, \mathrm{a}))$ (whose derivational complexity is 2). In other words, if conditions (i-v) are satisfied, it should follow that the description "$F_1(\mathrm{J}, \mathrm{a})$" actually means $J'a'$.

Since social norms can also determine an expression's meaning something, it will be helpful to introduce an indirection by stating that a conventional SCH-pattern is a stable linguistic use. This is what C-SLU does. In addition let us stipulate that a stable linguistic use of a description determines its meaning. This is what SLU-M does. C-M follows from their conjunction.

[11] The claim is supported by the argumentation that stable linguistic uses of expressions determine their meanings (§1.3).

C-SLU. If there is a conventional SCH-pattern for description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w, then there is a stable linguistic use of δ indicating μ with derivational complexity z among members of g at t in w.

SLU-M. If there is a stable linguistic use of description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among members of g at t in w.

Beneficiality of linguistic conventions For SCH-patterns to be a convention among a group of agents, it must be beneficial for them to do their part in it, that is, as a speaker, to utter a certain expression only if one has a certain attitude, and likewise as a hearer, to come to have a certain attitude when one hears a certain utterance. But how can a conventional SCH-pattern be beneficial for its members?

Suppose that there is a conventional SCH-pattern for the description of the name “a” between you and me. By being a convention, this entails that doing one’s part as a speaker or a hearer is beneficial for us. But can doing so be beneficial? It seems that only utterances that are performances of speech acts can be beneficial. But in ABLE-0, one cannot perform a speech act solely by uttering “a”. Such utterances would lack an illocutionary force.

What is missing to explain the beneficiality of “a” seems to be this. What can be used as an utterance part is beneficial because it makes systematic contributions to the SCH-patterns for descriptions of mood marked sentences. So, we can attribute values expressing the beneficiality to SCH-patterns for descriptions of mood marked sentences and then (try to) derive the respective values for the SCH-patterns for the descriptions of part-utterance types by working out what they contribute to the values of the SCH-patterns for descriptions of mood marked sentences.

9.3 Meaning in virtue of social norms

In this section, meaning in virtue of social norms is the topic. The approach mirrors the approach taken in the last section. To this end, I introduce *linguistic social norms*. In terms of this kind of social norms, the respective meaning-determination claim is stated. This is the first topic of this section. The two further topics are (i) the role of social structure and (ii) semantic normativity. Let me turn now to linguistic social norms.

The notion of a social norm is by itself not sufficient to state the meaning-determination claim for meaning in virtue of social norms. For social norms in general needn't have to do with language use at all. Also among the social norms that govern the use of expressions, not all have a meaning-constitutive role. There are, for example, social norms according to which one should not say that some product is "healthy" in advertisement or according to which one should not use taboo words like "fuck" or "nigger" on (American) television. Social norms of this kind just *regulate* how expressions already having a certain meaning are (not) to be used. According to social norms of this kind, the use of certain expressions is forbidden, optional, or required in certain situations. The characteristic patterns of such social norms are not SCH-patterns.

Contrast these social norms with the following cases: (i) the legislation of the meanings of trade names, (ii) the arbitration of meanings by experts, (iii) a group engaging in a normative debate about how to use and understand an expression. These cases involve a special kind of social norm which is constitutive for an expression's meaning something.

What is special about this second kind of social norms is that they concern the use and understanding of expressions. So, in line with the conventionalist part of my account, I suggest that linguistic social norms are a certain kind of social norms whose pattern of activity is an SCH-pattern. Besides having an SCH-pattern, linguistic social norms have a further feature: they relate to an additional group, the group of arbitrators. The role of the arbitrators is to select a particular SCH-pattern for an expression from the class of possible SCH-patterns for the expression. The selected SCH-pattern determines what it is to conform to and to deviate from it. In normal cases the enforcers are also the arbitrators. But for cases of division of linguistic labor the groups can be different. Let us thus define rationalistic linguistic social norms as follows:

There is a *rationalistic linguistic social norm to conform to the SCH-pattern Θ among the members of a group G (addressees) enforced by members of a group E (enforcers) accepting a system of norms N arbitrated by members of a group A (arbitrators) at time t in world w* iff

LSN1. there is a stable linguistic use among the arbitrators at time t in world w whose pattern of activity is Θ;

LSN2. there is a system of norms N according to which the addressees are N-required to conform to Θ and N-forbidden to deviate from Θ; and

LSN3. there is a rationalistic social norm to conform to Θ among the addressees enforced by the enforcers accepting N at time t in world w.

Clauses LSN1 and LSN2 incorporate the minimum we need to describe the role of the arbitrators, namely that it's their use and understanding that is enforced by the enforcers. Typically, arbitrators would or could correct other people's use and there might be a commonly known practice of deferral to them in case of dispute or uncertainty. But while these things are true of a typical scenario, it seems to me that their presence is not necessary for there to be a linguistic social norm. For this reason, I do not require that corrections or deferrals prevail among the respective groups.

The definition makes use of rationalistic social norms which are, in terms of rationality requirements, the most demanding kind of social norms (§8.2.2). In contrast, the weaker version of a rationally justifiable social norm does not require that the enforcers actually accept the relevant system of norms but only that they *could.* This weaker version can be used to define a corresponding notion of a *rationally justifiable* linguistic social norm that results from replacing "rationalistic" by "rationally justifiable" in the definition. I won't pursue this further and use "linguistic social norms" for both notions.

Having defined what a linguistic social norm is, we can state how an expression means something in virtue of a social norm. To this end, it is helpful to introduce a stipulation which "hides" certain details of the definiendum of a linguistic social norm. Let us say that "there is a linguistic social norm for description δ indicating content+ μ having a derivational complexity z addressed to members of group G at time t in world w" iff

1. there is an SCH-pattern Θ for δ indicating μ with derivational complexity z;
2. there is a group A of arbitrators such that there is a stable linguistic use to conform to Θ among its members at time t in world w;
3. there is a system of norms N according to which the addressees are N-required to conform to Θ and N-forbidden to deviate from Θ;
4. there is a group of enforcers E enforcing conformity to Θ and accepting systems of norms of which N is a part; and
5. there is a linguistic social norm to conform to the SCH-pattern Θ among the members of G enforced by members of E accepting a system of norms N arbitrated by members of A at time t in world w.

Now we claim that linguistic social norms are a kind of stable linguistic uses:

SN-SLU. If there is a linguistic social norm for description δ indicating content+ μ with derivational complexity z addressed to members of group g at time t in world w, then there is a stable linguistic use of δ indicating μ with derivational complexity z among members of g at t in w.

Together with the principle SLU-M that stable linguistic uses determine an expression's meaning something (§9.2), it follows that linguistic social norms determine an expression's meaning:

SN-M. If there is a linguistic social norm for description δ indicating content+ μ with derivational complexity z addressed to members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among the members of g at t in w.

9.3.1 The role of social structure

Social structure has different effects on the determination of an expression's meaning. I take the relevant social relations to be relations of power. For, if social structure has an influence on an expression's meaning something, then there is an inequality among us. Plausibly, this inequality is, at least in part, constituted by power. This is, in a way, the upshot of Humpty Dumpty's remark that the question is which is to be master. On this understanding, there are two kinds of roles of social structure:

- **Probabilistic roles**: (i) Making it likelier that an expression is used at all; (ii) Making it likelier that an expression is used in a certain group; (iii) Making it likelier that an expression is used uniformly; (iv) Making it likelier that an expression is used in a certain way
- **Meaning-determination role**: Determining an expression's meaning.

An example of a probabilistic role relates to social (network) structure. In societies that are socially structured, certain agents are more influential than others. Some of these agents are "hubs" which connect many agents with a small "social distance" (*e.g.* defined in terms of *being a friend of*). The use and understanding of an expression by a "hub" thereby tends to have more effect on the uses and understandings of the expression than the uses and understandings of a less influential agent.

I won't investigate the probabilistic roles here since I'm not discussing the evolution of stable linguistic use.[12] I'd like to illustrate the meaning-determination role by discussing two cases:

Case: expertise and expert power An interesting case is expertise and expert power, known from the debate about the division of linguistic labor. The thesis of the division of linguistic labor is a kind of social externalism about meaning: Facts about the social organization of the linguistic community determine, at least in part, what the expressions used by its members mean.[13]

The thesis of the division of linguistic labor seems to be a reasonable proposal about how the meaning of an expression is determined. To implement the proposal, a linguistic social norm would be helpful which gave experts a special linguistic right: the right to arbitrate semantic disputes about the term in question and, in particular, the right to decide what is the meaning of the expression in question.

If such a linguistic social norm exists, then it's the experts' use and understanding which selects the SCH-pattern for the expression in question (and thereby its meaning). It's the linguistic social norm which determines the expression's meaning by relating to the experts' SCH-pattern. In this scenario, the experts are the arbitrators. I'll call such linguistic social norms "expert social norms."

What is it to be an expert and how do expert social norms relate to French and Raven's *expert power*, which is the power an individual has in virtue of being skilled or expert about something over another persons who are in need of these skills or expertise (§8.1.1)?

Let's say that an expert about a word – say "cat" – is one who knows a lot about the word "cat" and is better at doing things with it. In a first approximation, this seems to be on the right track. Following Austin (1940) and Putnam (1997) we can describe the meaning of an expression by a vector consisting of syntactic information, sortal information, maybe a stereotype, maybe typical uses and inferential roles.[14] A (competent) language user

[12] But see for example the work of Buskens et al. (2008) and Corten and Buskens (2010) who show that network structure influences which equilibrium is reached. Thus, by controlling network structure or the relevant likelihoods, one can influence which linguistic conventions exist and thereby the meanings of the expressions in question.

[13] See *e.g* (Putnam 1997).

[14] If some of these things are part of an expression's meaning, then the semantics used by

knowing an expression's meaning, knows the information encoded in the vector. Importantly, Putnam's proposal does not entail that typical language users are fully informed about an expression's meaning. But experts about an expression, like our "cat" expert, know the expression's meaning in this sense. Moreover, a "cat" expert should be a reliable classifier of things into those to which the word "cat" applies and into those to which it doesn't apply. Since one can know more or less about something and be better or worse at doing things with it, being an expert is something gradable. This induces a comparative expertise relation among the members of some population.

From this it follows that by itself, expertise has no influence on an expression's meaning something. It's semantically inert because it's socially inert. For the relation is characterized in a way that does not imply that members of the community know who's an expert on what and whether there is a need of experts. Expert power in the sense of French and Raven adds these factors.

For a linguistic social norm, the existence of an expertise relation is beneficial. It provides a structure which can influence the interactions in the community, if (i) the linguistic social norm is sensitive to the level of expertise people have, (ii) people can recognize the level of expertise someone has.

It even seems that without (i) and (ii) satisfied, such social norms cannot be effective in the sense that they bring about the social arrangement that experts decide contested cases and that laymen defer their judgments to them and accept the expert decisions.

One way of satisfying (i) and (ii) is by a norm N that the enforcers of the respective expert social norm accept and that implements a sensitivity to expertise.[15] This relates to the information component required for expert power in the sense of French and Raven.

Often, there is also a need for expertise since we want experts to judge unsettled cases. Hence, it's also rational to want that experts have the right to arbitrate unsettled cases. This then brings French and Raven's *need* for expertise on the table which is a also condition for someone having expert power.

my account has to be revised.

[15] This is a way to understand so-called "dereferential conventions" that Blackburn (1984:130) discusses.

To conclude, the case of expert social norms illustrates some of the features of linguistic social norms. The experts are the arbitrators. They needn't be the enforcers but they are in a good position to be them since they know their stuff. Moreover, the addressees should accept the judgments of the experts.

Case: language rulers Another scenario illustrating linguistic social norms is the milk-scenario from the introduction in §1.3:

(10) In many countries there are laws – implemented as social norms – concerning the correct use of the trade name "milk;" violations are forbidden and can be punished. One can be demanded to use and understand the word accordingly. Different powers are involved, namely positional power and sanction power of the legislators (at least indirectly through the means of legal processes resulting in actions of the executive body). The addressees of the social norm are persons using the trade name "milk" for the purpose of doing business. The punishments (sanctions) create an incentive for them to conform to the regulations, that is, to apply "milk" only to milky liquids with at least 3% fat *etc.*

I've argued that (i) we're inclined to say that one of the meanings of "milk" is *milky liquid with at least* 3.0% *fat* in this scenario and that (ii) this is so in virtue of there being a linguistic social norm. Again, the meaning is selected by the arbitrators' SCH-pattern and determined by the linguistic social norm enforced by the executive body which is sensitive to the SCH-pattern selected by the arbitrators. There is a contrast with the expert social norm we've considered above. It's not necessarily the case that the experts are the arbitrators. Moreover, the scenario suggests that in case of a linguistic social norm for an expression, positional power lets one arbitrate what the expression means and sanction power lets one enforce conforming behavior. By these means, an expression can be "given" a meaning in a group.

Language rulings can be extreme as the ruling of the US Supreme Court in 1983 in the case Nix V Hedden illustrates (see Wikipedia 12.05.2010): The topic was whether tomatoes are fruits (as the scientific classification has it) or vegetables (for one had to pay a tax for imported vegetables but not for fruits). The court ruled that tomatoes are vegetables. The ruling can be described as one of selecting the meaning of "tomato."

9.3.2 Semantic normativity

On the basis of my account of meaning in virtue of social norms, I can now offer my final explanation of semantic normativity. In chapter 2 I argued that the characteristic feature of semantic normativity consists in being in a position to make normative demands:

N′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use e in accordance with its meaning m at $\mathcal{C}$ (by uttering "e means m at $\mathcal{C}$" which expresses an *ought* with a demanding character).

On the basis of the definition of a social norm and the analysis of the distinction between recommendations and demands (§8.2.1), we get the result that the enforcers of a social norm are in a position to make normative demands. Not all *oughts* that are expressed by an enforcer of a social norm are *semantic*. But the *oughts* that are expressed by an enforcer's demanding to conform to the SCH-pattern of a *linguistic social norm* are *semantic*. For linguistic social norms constitute (or determine) the meanings of expressions – or more precisely: descriptions:

N″. For all descriptions δ, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use δ in accordance with its meaning m at $\mathcal{C}$ (by uttering "δ means m at $\mathcal{C}$" which expresses an *ought* with a demanding character).

According to my proposal *to use δ in accordance with its meaning m* is to conform to the respective SCH-pattern Θ for δ.

An SCH-pattern describes the roles of both speakers and hearers (§9.1.2) and thereby determines for both speakers and hearers what it is to conform and what it is to deviate. For a speaker, to conform to an SCH-pattern is to only utter a certain expression if one is in a mental state whose content+ has a part that is indicated by the utterance. For a hearer, to conform to an SCH-pattern is to come to believe, upon hearing the utterance, that the speaker has a certain attitude which has a certain part that is indicated by the utterance.

Hence N″ entails that enforcers can demand speakers and hearers to conform and not to deviate from Θ. Consequently, the speakers and hearers among the addressees of a linguistic social norm ought to conform to (and not to deviate from) Θ.

For example, on the basis of a social norm, a speaker can be demanded not to utter an assertion-sentence unless she has the required belief. Likewise, a hearer can be demanded to come to believe that the speaker has the required belief.

Hence, speakers can be demanded to be truthful and hearers can be demanded to exhibit a certain linguistic understanding upon hearing an utterance.

9.4 Languages and Humpty Dumpty's problem

According to my account, expressions (or rather their descriptions) are not assigned a meaning relative to a language. The only use I make of languages is a limited technical use of formal languages for systematically describing expressions. I'll now relate meaningful expressions to notions such as *meaning something in a language*, *shared language*, and *public language*. Doing so helps us to make progress on the explanation of language varieties and the solution of Humpty Dumpty's problem. To this end, (i) I start by returning to the meaning-without-use problem. (ii) I provide definitions for several notions relating to languages. I indicate the roles they might play, in particular with respect to language varieties. (iii) I state my solution to Humpty Dumpty's problem.

Meaning without use The account developed so far also faces the meaning-without-use problem (§6.1.5): ABLE-0 consists of infinitely many expressions with compositional meanings. But only a part of ABLE-0 can be used by a population at any time, namely the part that is precisely defined by the prevailing stable linguistic uses with a certain derivational complexity. My proposal is as follows: The SCH-patterns of stable linguistic uses of descriptions or functions that prevail among a population are the patterns that the members of the population could realize. Thereby, also actually unrealized (parts of the) patterns can be the pattern of a stable linguistic use among the members of the population. Hence, the meaning-determination claims are so far stated for the *effective* stable linguistic uses among members of a population.

Effective stable linguistic uses for descriptions or functions cannot be used to explain the meanings of the expressions that couldn't be realized by the members of the respective population. To explain the meanings of these descriptions I reapply Schiffer's proposal for Lewis' theory: I stipulate that the SCH-pattern of a stable linguistic use is realized by a *translator* the members possess. I implement this idea below.

But in my theory, translators have a more moderate role: They are not needed for *any* expression – whether sentential or sub-sentential – in the effective language used by a population (and this fragment can be quite large since it also contains expressions that the members *would* use and understand in certain ways). The translators are only used to explain in virtue of what unusable meaningful expressions have their respective meanings.

Sociolects, shared languages, and total languages Let us define the *effective sociolect possessed by an agent a at a time t in a world w* as the set of all stable linguistic uses a is a member of, at t in w. The *effective shared language among members of a group g at a time t in a world w* is the intersection of the sociolects possessed by the members of g at t in w. The *effective total language among members of a group g at a time t in a world w* is the union of the sociolects possessed by the members of g at t in w.[16]

Languages in these senses are something social. They cannot be a private language or an idiolect that is solely possessed by an individual. For there is always a group among which a stable linguistic use prevails. Hence, I call that what is assigned to an agent a "sociolect" rather than an "idiolect." The other notions are defined in terms of sociolects. So, they inherit the social character.

Since my notions of languages are social, they do not fit a Chomskian understanding in terms of the "internalized" languages humans can possess. They also don't fit the typical approaches in sociolinguistics which base the classification of languages on the language attitudes "the folk" has.[17] For language attitudes are not used in my definitions above.[18]

Moreover, languages in these senses are not abstract languages in Lewis' sense (*i.e.* a mapping from forms into meanings). But we can define abstract

[16] The two definitions of languages in a group are inspired by Lewis' definitions in his article *Languages and language* (Lewis 1975:32).

[17] See (Garrett 2006) and (Niedzielski and Preston 2000).

[18] Language attitudes figure indirectly in the definitions. For in case of social norms, there are normative attitudes among the enforcers. According to them, one ought to use and understand expressions in accordance with their meanings.

languages for an agent a (and similarly for a group of agents) on the basis of the definitions provided above. We construct the effective abstract language $\mathcal{L}_{a,t,w}$ as the set of tuples $\langle \delta, \mu \rangle$ such that for some z and group g having a as a member, there is a stable linguistic use of description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w and δ's derivational complexity does not exceed z. (Observe the role of the derivational complexity parameter to define the effectively used language.)

On the basis of effective abstract languages, we can implement the proposal to deal with the meaning-without-use problem. The extended abstract language $\mathcal{L}_{\mathcal{T}a,t,w}$ of an agent a is based on her effective abstract language $\mathcal{L}_{a,t,w}$ as follows:

(11) Description δ means μ in $\mathcal{L}_{\mathcal{T}a,t,w}$ iff $\mathcal{L}_{a,t,w} \subseteq \mathcal{L}_{\mathcal{T}a,t,w}$, a processes $\mathcal{L}_{a,t,w}$-utterances via an $\mathcal{L}_{a,t,w}$-determining translator $\mathcal{T}$, $\mathcal{T}$ generates $\mathcal{L}_{\mathcal{T}a,t,w}$, and $\mathcal{T}$ pairs δ with μ.

The idea is that we extend effective abstract language on the basis of the translator that realizes a's stable linguistic use. It depends then on the properties of the translator whether the extended abstract language $\mathcal{L}_{\mathcal{T}a,t,w}$ is defined for infinitely many descriptions or not.

Another question that is still open and pertains to Humpty Dumpty's problem concerns the notion of a public language. I need the notion for two purposes: (i) to explicate the idiom "e means m in language $\mathcal{L}$" and (ii) to solve Humpty Dumpty's problem. For these purposes, I think, we can identify the notion of a public language with either the notion of a shared language or the notion of a total language. I choose the former option since the descriptions' meanings of a shared language are shared among the members of the group which is required if one wants to explain successful communication in terms of shared meanings.

Meaning something in a language On the basis of the proposal advocated here, we can reconstruct what seems to be the main role of the idiom "e means m in language $\mathcal{L}$", namely to distinguish between homonyms which belong to different languages. For example, while the word spelled as "rot" names a color in German, what is so spelled has different meanings in English, among them being *decay*.

Since I individuate expressions historically (as Millikan and lexicographers do), the way an expression is spelled or pronounced is not decisive for saying whether two expressions are identical or not. For example, "rot" in German and "rot" in English would already come out as two distinct expressions on my proposal since they belong to different histories. That is, by individuating expressions differently, the need to assign meanings relative to a language is less obvious. For this reason, it seems to me to be justified that in my account, descriptions are assigned meanings not relative to a language.

Nevertheless, we can give an analysis of the idiom "e means m in language $\mathcal{L}$." My proposal is as follows: (i) e is a name for a sound or spelling type, (ii) $\mathcal{L}$ is a name for a shared language, (iii) "in $\mathcal{L}$" is a modifier of e and yields a definite description for an expression: the expression that is of the sound or spelling type e and that is part of $\mathcal{L}$. Within a community of communicating agents, we expect that most spelling (or sound) types just have one meaning, unless there are features in their SCH-patterns that allow the speakers and hearers to tell the two uses and understandings apart (§7.3.2.4). Consequently, we should expect that on my pragmatic analysis one can uniquely pick out an expression by using the idiom "e ... in language $\mathcal{L}$."

Humpty Dumpty's problem Finally I can state my solution of Humpty Dumpty's problem (§1.4):

HD. If it's the use of expressions in a community which determines their meanings, then in which way does this determination depend on the members and the circumstances?

If one endorses my account, there is an obvious answer to HD: In case of a linguistic (normative) convention for an expression having a certain description, the description's having a certain meaning depends on *all* members of the convention. In case of a linguistic social norm for an expression, it depends on the enforcers and arbitrators what the expression means. In extreme cases, this can be a single person. Hence, if Humpty Dumpty is "the master" – that is, if he is the sole arbitrator or enforcer –, then he is right. But it seems that this wouldn't be the normal case which seems to be more egalitarian.[19]

[19] Social structure also plays a probabilistic role in the determination of meaning. Thus, a more complete solution of the problem should also include this, especially when it comes

9.5 Evaluation

In my account descriptions of expressions are assigned structured mental contents+ as meanings. The assignments are determined by stable linguistic uses (linguistic conventions, social norms, and normative conventions). For each description of a word, there is a separate stable linguistic use. This contrasts with other conventionalist accounts. Thereby, language varieties can be explained in a better way. For there is no need to assign possibly big populations whole languages. Rather, the languages of a group are defined by the stable linguistic uses prevailing among the members of the group.

I illustrated the account on the basis of the "toy" language ABLE-0. The language is severely restricted: (i) There are its expressive limits for there is no quantification and there are no adverbials, indexicals, and vague expressions. The lack of these features severely limits the explanatory power of my account. (ii) The semantics is simplistic and naive. Moreover, if one endorses the thesis of the division of linguistic labor (§9.3.1), then the semantic theory has to be changed to accommodate the proposed more complex meanings (*e.g.* including stereotypes). To address these limitations, more complex SCH-patterns have to be defined and a state-of-the-art semantics has to be used. Such an extension would naturally be supplemented by a formalization of the core account; one could use an event calculus to define SCH-patterns and a first-order logic to systematize the derivations of meanings.

Nevertheless, ABLE-0 illustrates how one can deal with structured expressions.

The "meanings" assigned to descriptions are considered to be *literal meanings.* But I've said nothing about their relation to what-is-said and other pragmatic notions of meaning – like metaphoric meanings or implicatures. Clearly, one should make these relations explicit by using a state-of-the-art pragmatic theory (which includes a theory of linguistic communication). My account is committed to the claim that such a theory is coherent with the one that is implicitly used by my way of describing linguistic transactions in terms of SCH-patterns. This rules out Gricean proposals in terms of intention-recognition. But as far as I can see, local changes to the SCH-patterns are sufficient to create a version of my proposal which is coherent

to the "circumstances" in HD. Since this would touch topics on the evolution of (linguistic) behavior in groups, I do not make it a topic here. But check the exemplary references I provide in §9.3.1 on the role of social network structure.

with other pragmatic theories. I think the changes can be limited to elements concerning the semantic values assigned to mood marked sentences (like "A:s"); in particular: the description of the S- and H-parts of the SCH-patterns that relate to the semantic values – and if it simplifies things, the semantic values themselves.

Even without having said which theory of linguistic communication I use, I can make some claims about Davidsonian uses (uses of malapropisms and the like) based on my account. Consider again Mrs Malaprop's use of "a nice derangement of epitaphs" when she should have used "a nice arrangement of epithets" (§3.1.3). If her use is not in accordance with a stable linguistic use, then on the basis of my account, her expression does not mean *a nice arrangement of epithets*.

This does not rule out that communication can be successful if someone uses the expression as Mrs Malaprop does. However, communication is then not successful because the meaning of her expression is determined by a stable linguistic use (and thereby shared). Rather, it is successful because the hearer is able to recover the intended meaning of the expression. I don't explain how such recoveries work. In §3.4 I proposed that a theory of linguistic communication should explain this.

Last but not least there is a worry that the use of categorial grammar to describe the language use that could prevail in groups is less innocent than I've claimed. The worry is that somehow, syntax might be *prior* (in some sense) to semantics. One reason to think so is that, due to the category-type homomorphism of the categorial grammar, the syntactic description already encodes semantic information. Another reason is that syntactic complexity is defined in terms of it which is, as I've said, implausible. I leave this worry as an open question for future research.

Adequacy Let us go through the adequacy conditions of §1.4 to assess my account. The verdict is that it does quite well but it does not explain the dynamics and evolution of language use.

I use Millikan's account of conventions. So I can refer to the evaluation in §7.1.5, which was positive. Millikan's account is adequate except for the following: the dynamics of conventions is not explained since there is no formal theory. There is an open problem with respect to the application of Millikan's account to my account: To show that the SCH-patterns of linguistic conventions are beneficial for their members, a value has to be derived that expresses the beneficiality of the SCH-pattern.

The result is similar for social norms (§8.3): The account is now adequate except for the explanation of the dynamics of social norms that is missing. In this chapter, I've shown how the role of arbitrators can be included (§9.3). Thereby, my explication of social norms is faithful to the pre-theoretic characterization of them (§1.2).

Let us now turn to the desiderata for the account of meaning provided in this chapter:

DesM1. The account must be coherent with an explanation of the communicative functions of meaning-sentences.

On the basis of my argumentation about semantic normativity (chapter 2), my account of social norms (chapter 8), and its application in this chapter (§9.3.2), the communicative functions can be explained. In particular, the demanding character of the *oughts* that can be expressed by uttering a meaning sentence has been explained in detail.

DesM2. The account must explain the meanings of expressions in terms of their stable uses (conventions, social norms, and normative conventions).

My account implements the explanatory architecture advocated in §7.3.1. Thereby, DesM2 is satisfied for there is meaning in virtue of all kinds of linguistic stable uses.

DesM3. The account must provide a plausible notion of a semantic mistake.

This desideratum is satisfied by my way of satisfying DesM1 (see above). I've argued for a notion of a semantic mistake in terms of linguistic social norms in §2.2.3.

DesM4. The account must allow for a plausible conception of a public language.

Since public languages are explained on the basis of expressions that already mean something among members of a group, language varieties can be explained in a better way (§9.4).

DesM5. The account must explain the usual meaning facts.

Since ABLE-0 and the SCH-patterns for it are designed so as to allow for structured expressions that have compositional meanings, the usual meaning facts (such as: an expression means something at all, there are systematic semantic relations between expressions) can be explained.

DesM6. The account must provide a solution to Humpty Dumpty's problem.

On the basis of my notion of a public language and the distinction between conventions and social norms (§1.2), I can give a plausible solution to Humpty Dumpty's problem (§9.4).

9.6 Summary

In this chapter I developed an alternative conventionalist account. It uses Millikan's notion of a convention, my account of social norms, and a new description of the communicative patterns in terms of SCH-patterns. The account is adequate except for the dynamics of language use, the mentioned restrictions due to the expressive limits of ABLE-0, and the worry that syntax is prior to semantics.

In comparison to the rival accounts I've considered, my account uses a distinguishing combination of features that make it better: (i) The account is non-Gricean and non-Intellectual. (ii) The account is compatible with dispositional conventions. (iii) The account can deal well with language varieties.[20] (iv) The account can deal well with the meanings of sub-sentential expressions (thereby facing less indeterminacy).[21] (v) The account can deal well with the meaning-without-use problem.[22] (vi) The account can explain meaning in virtue of social norms. (vii) The account can explain semantic normativity. (viii) The account provides a plausible solution to Humpty Dumpty's problem.

[20] The first three features are shared with Millikan's account; see §7.1.

[21] This feature is shared with Evolutionary Signaling Games theories that use structured meanings and signals; see §7.2.3.

[22] This feature is shared with my understanding of Actual Language Relation theories (§6.1.5) and my understanding of Evolutionary Signaling Games theories that use structured meanings and signals (§7.2.3).

Chapter 10

Conclusions & Outlook

> If I was a woman, and you were a man
> You would do things that I don't understand
> You say "Boo Boo Boo", and I think "La La La"
> Maybe "Boo" will mean "Ja" or just "Blah Blah Blah"
>
> *3 minutes in the brain of Bonaparte*
> Bonaparte

In the preceding chapters, I've discussed key topics of the conventionalist project: conventions, social norms, how they might determine meaning, and semantic normativity. Conventionalists face many powerful objections. But they've developed clever rejoinders, rebutting not only (the philosopher) Davidson but also (the musician) Bonaparte insofar as his remark was meant to be about literal meaning. To end this thesis, I'd like (i) to establish the dialectic situation, (ii) to highlight some insights, and (iii) to point to some open tasks for future research.

The dialectic situation On the basis of my argumentation, I've submitted the following theses:

C″. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable linguistic use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$.

N′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand from the addressees to use e in accordance with its meaning m at $\mathcal{C}$ (by uttering "e means m at $\mathcal{C}$" which expresses an *ought* with a demanding character).

The conventionality thesis C'' resulted from two revisions – one with respect to meaning in virtue of social norms (§1.3 where I introduced "stable uses" as a cover term for (normative) conventions and social norms), the other effected by the meaning-without-use problem (§6.1.5 where I introduced "stable linguistic uses" as a cover term for linguistic (normative) conventions and social norms whose behavioral dispositions are, in part, brought about by a translator). The normativity thesis N' resulted from one revision due to meaning in virtue of conventions (§2.4).

Together they form a powerful adequacy constraint: no conventionalist account in the literature entails both of them (to the best of my knowledge).

This fact notwithstanding, the accounts in this thesis are often much more defensible than people have thought. I hope to have made a convincing case for Lewis' proposals. In particular, the common Intellectuality objection is much too quick (unless Lewisians insist on "being Gricean," *i.e.*, if they insist on endorsing the conditional: if a speaker utters something conforming to a linguistic convention, then she means something – in Paul Grice's sense of "speaker-meaning"; see §5.3.4 and §6.2.4). The needed rationality assumptions are not that strong after all. But even a careful defense seems to hit a limit: Lewis' Actual Language Relation theory explains neither meaning in virtue of social norms nor semantic normativity. Its application to language-varieties cases is discouragingly complicated. The sentential primacy thesis effects underdetermined sub-sentential meanings. Some of these issues might be addressed if one developed Wayne Davis' account further.

I also hope to have shown that evolutionary accounts do not magically address all the problems. Millikan's account also has problems with sub-sentential expressions and, again, there is neither meaning in virtue of social norms nor an explanation of semantic normativity.

Huttegger's partial implementation of Millikan's account improves on Millikan. Together with the addition of syntactic structure to signaling games by Nowak *et al.*, I think that the paradigm of Evolutionary Signaling Games has a promising outlook. But in its present state, there is no meaning in virtue of social norms and no complete explanation of semantic normativity.

My account addresses the above points but so far it is only developed for a tiny fragment of language. Moreover, it lacks a formal model to explain the dynamics and evolution of language (use). Thereby, I cannot explain with rigor why it is beneficial to have a convention for a word. Last but not least, there is the worry that syntax might be prior to semantics.

Insights Looking back, I think the following insights were productive for carrying out the conventionalist project:

- The meaning-without-use problem: It made us aware that we need to add a cognitive component to the meaning-determination claims. My favored approach is based on Schiffer's $\mathcal{L}$-determining translator proposal.
- Rationality assumptions: Questioning the rationality assumptions resulted in a rethinking of central assumptions of conventionalist accounts, in particular of the description of communication in Gricean terms and of the conception of a convention. Millikan showed us how to think differently.
- Sub-sentential meanings: Focusing not only on sentential meanings but also on sub-sentential meanings, as Davis advised, uncovered gaps and issues with present accounts.
- Semantic normativity as part of the conventionalist project: Studying semantic normativity from the perspective of a foundational theory of meaning informed the debate. The principle I advocate is: *If meaning is normative, then this is because of the way it is determined* (MD-ME).
- Davidsonian uses (like malapropisms): Davidsonian uses of language make us aware of open questions of the conventionalist project. The interaction between the core theories – a foundational theory of meaning, a semantic theory, and a pragmatic theory – is still to be explored.

Outlook If my argumentation was so far by and large correct, then the following accounts are viable:

(i) Evolutionary Signaling Games accounts: The open tasks include the study of the philosophical foundations of the evolution-theoretic foundations, the inclusion of social norms as a notion that is distinct from conventions, the study of signaling games with infinitely many signals, and relating such accounts to a pragmatic theory.

(ii) Actual Language Relation accounts: The most promising account seems to me to be Davis'. The open tasks include: addressing the meaning-without-use problem, explaining language varieties, and separating conventions from social norms.

(iii) My account: The open tasks include the extension of ABLE-0 to cover a larger fragment of language using a state-of-the-art semantics, the development of a formal model to explain the dynamics and evolution of language (use), addressing the “syntax is prior to semantics” worry, and relating the account to a pragmatic theory.

In connection with pragmatic theories, it will be of particular interest to see how communication systems can evolve if the communicators use and understand utterances according to a particular pragmatic theory (§1.1.1; *e.g.* Sperber and Wilson’s relevance theory (Sperber and Wilson 1995), Borg’s minimal semantics (Borg 2006), Recanati’s contextualism (Recanati 2004), and Franke’s iterated best response model (Franke 2009) – which is interesting since it uses signaling games with meaningful signals). For language is often used to bring about perlocutionary effects that, in a sense, go beyond understanding the literal meanings of the uttered expressions. How can an agent not knowing the literal meanings of expressions learn them if the reaction the speaker wants her to show is a certain perlocutionary effect? Another way of describing the problem is this: Often what a speaker meant is not what her words literally mean. The speakers often rely on the hearers’ figuring out what they meant on the basis of their understanding of the expression’s literal meaning. The hearers, in turn should understand what the respective speakers meant, at least often enough. In such instances of linguistic communication pragmatic reasoning plays an important role. The problem to solve is how one can determine the literal meanings of the expressions on the basis of such instances of linguistic communication.

Bibliography

Alexander, J. McKenzie. 2009. "Evolutionary game theory." In Zalta (2009).

Aumann, Robert J. 1976. "Agreeing to disagree." *The annals of statistics* 4:1236–1239.

Austin, John L. 1940. "The meaning of a word." In James O. Urmson and Geoffrey J. Warnock (eds.), *Philosophical Papers*, 55–75. Oxford [u.a.]: Oxford University Press, 3 edition.

—. 1975. *How to do things with words (William James Lectures).* Cambridge: Harvard University Press, 2 edition.

Axelrod, Robert. 2006. *The evolution of cooperation.* New York, NY: Basic Books, 2 edition.

Aydede, Murat. 2008. "The language of thought hypothesis." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition).*

Bach, Kent and Harnish, Robert M. 1980. *Linguistic communication and speech acts.* Cambridge, MA; London: MIT Press.

Barwise, Jon and Perry, John. 1986. *Situations and attitudes.* Cambridge, MA: Bradford Book.

Bays, Timothy. 2001. "On Putnam and his models." *Journal of Philosophy* 98:331–350.

—. 2009. "Skolem's paradox." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition).*

Bennett, Jonathan Francis. 1973. "The meaning-nominalist strategy." *Foundations of Language* 10:141–168.

—. 1976. *Linguistic behaviour.* Cambridge: Cambridge University Press.

Benz, Anton, Jäger, Gerhard, and van Rooij, Robert. 2005a. "An introduction to game theory for linguists." In Benz et al. (2005b).

Benz, Anton, Jäger, Gerhard, and van Rooij, Robert (eds.). 2005b. *Game theory and pragmatics.* Basingstoke, Hampshire [u.a.]: Palgrave MacMillan.

Bicchieri, Cristina. 2006. *The grammar of society: The nature and dynamics of social norms.* Cambridge: Cambridge University Press.

Bilgrami, Akeel. 1993. "Norms and meaning." In Stoecker (1993), 121–144.

Binmore, Ken G. 2008. "Do conventions need to be common knowledge?" *Topoi* 27:17–27.

Blackburn, Simon. 1984. *Spreading the word: Groundings in the philosophy of language.* Oxford: Clarendon Press.

Boer, Bart de and Zuidema, Willem H. 2009. "Models of language evolution: Does the math add up?" Amsterdam.

Boghossian, Paul A. 1989. "The rule-following considerations." *Mind* 98:507–549.

—. 2003. "The normativity of content." In Ernest Sosa and Enrique Villanueva (eds.), *Philosophy of mind*, volume 13 of *Philosophical issues*, 31–45. Boston: Blackwell.

—. 2005. "Is meaning normative?" In Nimtz and Beckermann (2005), 205–218.

Bonaparte. 2008. "Too much." CD.

Borg, Emma. 2006. *Minimal semantics.* Oxford: Oxford University Press.

Boyd, Robert and Richerson, Peter J. 1992. "Punishment allows the evolution of cooperation (or anything else) in sizable groups." *Ethology and Sociobiology* 13.

Brandom, Robert B. 1994. *Making it explicit: Reasoning, representing, and discursive commitment.* Cambridge, MA: Harvard University Press.

Burge, Tyler. 1975. "On knowledge and convention." *The Philosophical Review* 84:249–255.

Burke, Mary A. and Young, H. Peyton. 2009. "Social norms: Forthcoming in The Handbook of Social Economics, edited by Alberto Bisin, Jess Benhabib, and Matthew Jackson. Amsterdam: North-Holland."

Buskens, Vincent, Corten, Rense, Weesie, and Jeroen. 2008. “Consent or conflict: coevolution of coordination and networks.” *Journal of Peace Research* 45:205–222.

Camerer, Colin F. 2003. *Behavioral game theory: experiments in strategic interaction.* New York [u.a.]: Russell Sage Foundation.

Camus, Albert. 1984. *Caligula*, volume 4 of *Cahiers Albert Camus.* Paris: Gallimard.

Carroll, Lewis. 1994. *Through the looking glass.* London: Penguin.

Chalmers, David J. and Jackson, Frank. 2001. “Conceptual analysis and reductive explanation.” *The Philosophical Review* 110:315–360.

Cheney, Dorothy L. and Seyfarth, Robert M. 1992. *How monkeys see the world: Inside the mind of another species.* Chicago, IL: University of Chicago Press, 1 edition.

Chomsky, Noam. 1980. *Rules and representations*, volume 11 of *Woodbridge lectures.* New York: Columbia University Press.

Cohen, Philip R. and Levesque, Hector J. 1990. “Intention is choice with commitment.” *Artificial Intelligence* 42:213–261.

Collins, Michael P. 2009. *The nature and implementation of representation in biological systems.* Ph.D. thesis, City University of New York, New York.

Corten, Rense and Buskens, Vincent. 2010. “Co-evolution of conventions and networks: An experimental study.” *Social Networks* 32:4–15.

Crawford, Vincent P. and Sobel, Joel. 1982. “Strategic information transmission.” *Econometrica* 50:1431–1451.

Cubitt, Robin P. and Sugden, Robert. 2003. “Common knowledge, salience and convention: A reconstruction of David Lewis’ game theory.” *Economics and Philosophy* 19:175–210.

Darwin, Charles R. 1876. *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London: John Murray, 6 edition.

Davidson, Donald. 1978. “What metaphors mean.” *Critical Inquiry* 5:31–47.

—. 1993. “Reply to Akeel Bilgrami.” In Stoecker (1993), 145–147.

—. 2001. "Communication and convention." In Donald Davidson (ed.), *Inquiries into truth and interpretation*, 265–280. Oxford: Clarendon Press.

—. 2005. "A nice derangement of epitaphs." In *Truth, language, and history*, 89–107. Oxford: Clarendon Press.

Davis, Wayne A. 2003. *Meaning, expression, and thought.* Cambridge studies in philosophy. Cambridge: Cambridge University Press.

—. 2005. *Nondescriptive meaning and reference: An ideational semantics.* Oxford: Clarendon Press.

Dowty, David R., Wall, Robert E., and Peters, Stanley. 1992. *Introduction to Montague semantics*, volume 11 of *Studies in Linguistics and Philosophy.* Dordrecht: Kluwer Academic Publishers.

Emerson, Richard M. 1962. "Power-dependence relations." *American Sociological Review* 27:31–41.

Evans, Gareth and McDowell, John H. (eds.). 1976. *Truth and meaning: Essays in semantics.* Oxford: Clarendon Press.

Fiske, Susan T. and Taylor, Shelley E. 1991. *Social cognition.* MacGraw-Hill series in social psychology. New York NY: MacGraw-Hill.

Fodor, Jerry A. 1987. *Psychosemantics: The problem of meaning in the philosophy of mind*, volume 2 of *Bradford books.* Cambridge, MA: MIT Press.

Fodor, Jerry A. and Lepore, Ernest. 1992. *Holism: A shopper's guide.* Oxford: Blackwell.

Frank, Robert H. 1988. *Passions within reason: The strategic role of the emotions.* New York, NY: Norton.

Franke, Michael. 2009. *Signal to act.* Ph.D. thesis, Universiteit van Amsterdam, Amsterdam.

Frankfurt, Harry G. 1971. "Freedom of the will and the concept of a person." *The Journal of Philosopy* 68:5–20.

French, John R. P. and Raven, Bertram. 1960. "The bases of social power." In Dorwin Cartwright and Alvin Zander (eds.), *Group dynamics*, 607–623. London: Tavistock Publications, 2 edition.

Gärdenfors, Peter. 1993. "The emergence of meaning." *Linguistics and Philosophy* 16:285–309.

Garrett, Peter. 2006. “Language attitudes.” In Carmen Llamas, Louise Mullany, and Peter Stockwell (eds.), *The Routledge companion to sociolinguistics*, Routledge Companions, 116–121. London: Routledge.

Gauker, Christopher. 2005. “Review of Meaning, Expression and Thought: Wayne A. Davis, Cambridge University Press, 2003, xvii + 654 pp.” *Journal of Pragmatics* 37:955–959.

George, Alexander. 1990. “Whose language is it anyway.” *The Philosophical Quarterly* 40:275–298.

Gerbrandy, Jelle. 1999. *Bisimulations on planet Kripke.* Ph.D. thesis, Universiteit van Amsterdam, Amsterdam.

Gibbard, Allan. 1990. *Wise choices, apt feelings: A theory of normative judgment.* Oxford: Clarendon Press.

—. 1994. “Meaning and normativity.” *Philosophical Issues* 5:95–115.

—. 2003. *Thinking how to live.* Cambridge, MA [u.a.]: Harvard University Press.

Gilbert, Margaret. 1983. “Agreements, conventions, and language.” *Synthese* 54:375–407.

—. 1989. *On social facts.* International library of philosophy. London: Routledge.

—. 2008. “Social conventions revisited.” *Topoi* 27:5–16.

Glock, Hans-Johann. 2005. “The normativity of meaning made simple.” In Nimtz and Beckermann (2005), 219–241.

—. 2010. “Does language require conventions?” In Pasquale Frascolla, Diego Marconi, and Alberto Voltolini (eds.), *Wittgenstein.* Basingstoke: Palgrave MacMillan.

Glüer, Kathrin and Wikforss, Åsa M. 2008. “Against normativity again: reply to Whiting.”

—. 2009. “The normativity of meaning and content.” In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition).*

Grafen, Alan. 1990. “Biological signals as handicaps.” *Journal of Theoretical Biology* 144:517–546.

Grandy, Richard E. 1977. “Review of D. Lewis: Convention.” *Journal of Philosophy* 74:129–139.

Grice, Herbert Paul. 1957. “Meaning.” *The Philosophical Review* 66:377–388.

—. 1969. “Utterer’s meaning and intentions.” *The Philosophical Review* 78:147–177.

—. 1989a. “Logic and conversation.” In Grice (1989c), 22–46.

—. 1989b. “Meaning revisited.” In Grice (1989c), 283–303.

Grice, Herbert Paul (ed.). 1989c. *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Grice, Herbert Paul. 1989d. “Utterer’s meaning, sentence meaning, and word-meaning.” In Grice (1989c), 117–137.

Guldborg Hansen, Pelle. 2009. *Social convention: In defense of David Lewis.* Ph.D. thesis, Roskilde Universitet, Roskilde.

Hájek, Alan. 2007. “David Lewis.” In Noretta Koertge (ed.), *New dictionary of scientific biography*, volume 4. Detroit: Charles Scribner’s Sons.

Hanks, Patrick. 2009. “Review: Stephen J. Perrault (ed.). Merriam-Webster’s Advanced Learner’s English Dictionary.” *International Journal of Lexicography* 22:301–315.

Harms, William F. 2004. *Information and meaning in evolutionary processes.* Cambridge studies in philosophy and biology. Cambridge: Cambridge University Press.

Hart, Herbert L. A. 1997. *The concept of law.* Oxford [u.a.]: Clarendon Press, 2 edition.

Hartogh, Govert A. den. 2002. *Mutual expectations: A conventionalist theory of law*, volume 56 of *Law and philosophy library.* The Hague: Kluwer Law International.

Hattiangadi, Anandi. 2006. “Is meaning normative?” *Mind & Language* 21:220–240.

—. 2007. *Oughts and thoughts: Rule-following and the normativity of content.* Oxford: Clarendon Press.

Hendriks, Herman. 2001. “Compositionality and model-theoretic interpretation.” *Journal of Logic, Language and Information* 10:29–48.

Henrich, Joseph, Heine, Steven J., and Norenzayan, Ara. forthcoming. “The weirdest people in the world?” *Behavioral and Brain Sciences* .

Hofbauer, Josef and Sigmund, Karl. 2003. "Evolutionary game dynamics." *Bulletin of the American Mathematical Society* 40:479–519.

Hopcroft, John E., Motwani, Rajeev, and Ullman, Jeffrey D. 1979. *Introduction to automata theory, languages, and computation.* Reading, MA [u.a]: Addision-Wesley.

Horowitz, Damon. 2007. "Davidson's intentions and Reimer's conventions."

Horwich, Paul. 2004. *Meaning.* Oxford: Clarendon Press.

Hume, David. 2003. *A treatise of human nature.* Mineola N.Y.: Dover Publications.

Hunter, David. 2003. "Is thinking an action?" *Phenomenology and the Cognitive Sciences* 2:133–148.

Huttegger, Simon M. 2007. "Zur Evolution von Normen." In Günther Kreuzbauer, Norbert Gratzl, and Ewald Hiebl (eds.), *Persuasion und Wissenschaft*, volume 2 of *Salzburger Beiträge zu Rhetorik und Argumentationstheorie*, 267–277. Wien: Lit.

Itkonen, Esa. 2008. "The central role of normativity in language." In Jordan Zlatev, Timothy P. Racine, Chris Sinha, and Esa Itkonen (eds.), *The shared mind*, volume 12 of *Converging Evidence in Language and Communication Research*, 279–305. Amsterdam: Benjamins.

Jackman, Henry. 1998. "Convention and language." *Synthese* 117:295–312.

Jackson, Frank. 1998. *From metaphysics to ethics.* Oxford: Clarendon Press.

Jamieson, Dale. 1975. "David Lewis on Convention." *Canadian Journal of Philosophy* 5:73–81.

Janssen, Theo M. V. 1997. "Compositionality." In Johan van Benthem and Alice ter Meulen (eds.), *Handbook of logic and language*, 417–473. Amsterdam [u.a.]: Elsevier.

Jeffrey, Richard C. 1990. *The logic of decision.* Chichago: University of Chicago Press, 2 edition.

—. 2004. *Subjective probability: The real thing.* Cambridge: Cambridge University Press.

Jorgensen, Andrew K. 2008. "Lewis's synthesis." *International Journal of Philsophical Studies* 16:77–84.

Kant, Immanuel. 1968. *Kritik der praktischen Vernunft*, volume 5 of *Kants Werke, Akademie Textausgabe.* Berlin: Walter de Gruyter.

Kaplan, David. 1990. "Words." *The Aristotelian Society supplementary volume* 64:93–119.

Kemmerling, Andreas. 1976. *Konvention und sprachliche Kommunikation.* Ph.D. thesis, Ludwig Maximilians Universität, München.

—. 1979. "Was Grice mit "Meinen" meint." In Günther Grewendorf (ed.), *Sprechakttheorie und Semantik*, volume 276 of *Suhrkamp-Taschenbuch Wissenschaft*, 67–118. Frankfurt am Main: Suhrkamp, 1 edition.

—. 1986. "Utterer's meaning revisited." In Richard E. Grandy and Richard Warner (eds.), *Philosophical grounds of rationality*, Clarendon paperbacks, 131–155. Oxford: Clarendon Press, 1 edition.

—. 1993. "The philosophical significance of a shared language." In Stoecker (1993), 85–116.

—. 1997. "Der bedeutungstheoretisch springende Punkt sprachlicher Verständigung." In Geert-Lueke Lueken (ed.), *Kommunikationsversuche*, volume 7 of *Leipziger Schriften zur Philosophie*, 60–106. Leipzig: Leipziger Universitätsverlag.

Kim, Jaegwon. 1998. *Philosophy of mind.* Dimensions of philosophy series. Colorado, CO: Westview Press.

Kintisch, Eli. 30.11.1999. "New Jewish Rep. makes splash." *JTA* .

Knight, Jack. 2004. *Institutions and social conflict.* The political economy of institutions and decisions. Cambridge: Cambridge University Press.

Kölbel, Max. 1998. "Lewis, language, lust and lies." *Inquiry* 41:301–315.

—. 2001. "Two dogmas of Davidsonian semantics." *The Journal of Philosopy* 98:613–635.

Kripke, Saul A. 2000. *Wittgenstein on rules and private language: An elementary exposition.* Cambridge: Harvard University Press.

Kuhn, Steven. 2007. "Prisoner's dilemma." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2007 Edition).*

Larson, Richard K. and Segal, Gabriel. 1995. *Knowledge of meaning: An introduction to semantic theory.* A Bradford book. Cambridge, MA: MIT Press.

Laurence, Stephen. 1996. "A Chomskian alternative to convention-based semantics." *Mind* 105:269–301.

Lenaerts, Tom and Vylder, Bart de. 2005. "On the evolutionary dynamics of meaning-word associations." In Benz et al. (2005b), 263–284.

Lepore, Ernest and Ludwig, Kirk. 2007. *Donald Davidson: Meaning, truth, language, and reality.* Oxford: Oxford University Press.

Lewis, David K. 1970. "General semantics." *Synthese* 22:18–67.

—. 1975. "Languages and language." In Keith Gunderson (ed.), *Language, mind and knowledge*, 3–35. Minneapolis: University of Minnesota Press.

—. 1976. "Convention: A reply to Jamieson." *Canadian Journal of Philosophy* 6:113–120.

—. 1978. "Truth in fiction." *American Philosophical Quarterly* 15:37–46.

—. 1992. "Meaning without use: Reply to Hawthorne." *Australasian Journal of Philosophy* 70:106–110.

—. 2002. *Convention: a philosophical study.* Oxford [u.a.]: Blackwell.

Lieberman, Mark. 26.05.2006. "Monkey words."

Loar, Brian. 1976. "Two theories of meaning." In Evans and McDowell (1976).

Locke, John. 1970. *Two treatises of government: A critical edition with an introduction and appartus criticus by Peter Laslett.* Cambridge: Cambridge University Press, 2 edition.

Lycan, William G. 2004. *Philosophy of language: a contemporary introduction: Routledge contemporary introductions to philosophy.* London [u.a.]: Routledge & Kegan Paul.

—. 2008. *Philosophy of language: a contemporary introduction.* New York [u.a.]: Routledge, 2 edition.

Magen, Stefan. 2005. "Fairness, Eigennutz und die Rolle des Rechts: Eine Analyse auf Grundlage der Verhaltensökonomik." Bonn.

Marmor, Andrei. 2009. *Social conventions: From language to law.* Princeton monographs in philosophy. Princeton, N.J: Princeton University Press.

Maynard Smith, John and Price, George R. 1973. "The logic of animal conflict." *Nature* 246:15–18.

Merriam Webster Online Dictionary. 2010. "”camera”."

Millar, Alan. 2002. "The normativity of meaning." In Anthony O'Hear (ed.), *Logic, thought, and language*, volume 51 of *Royal Institute of Philosophy supplement*, 57–73. Cambridge: Cambridge University Press.

—. 2004. *Understanding people: Normativity and rationalizing explanation.* Oxford, New York: Clarendon Press.

Millikan, Ruth G. 1990. "Truth rules, hoverflies, and the Kripke-Wittgenstein paradox." *The Philosophical Review* 99:323–353.

—. 1995a. *Language, thought, and other biological categories: New foundations for realism.* Cambridge: MIT Press.

—. 1995b. "Pushmi-pullyu representations." *Philosophical Perspectives* 9:185–200.

—. 1998. "Language conventions made simple." *Journal of Philosophy* 95:161–180.

—. 2005a. "In defense of public language." In Millikan (2005b), 24–52.

—. 2005b. *Language: A biological model.* Oxford: Clarendon Press.

—. 2005c. "On meaning, meaning and meaning." In Millikan (2005b), 53–76.

—. 2005d. "Proper function and convention in speech acts." In Millikan (2005b), 139–165.

—. 2005e. "The language-thought partnership: A bird's eye view." In Millikan (2005b), 92–105.

—. 2006. *Varieties of meaning: The 2002 Jean Nicod lectures.* London: Bradford Books.

—. 2008. "A difference of some consequence between conventions and rules." *Topoi* 27:87–99.

Millikan, Ruth Garrett. 2000. *On clear and confused ideas: An essay about substance concepts.* Cambridge studies in philosophy. Cambridge, UK: Cambridge University Press.

Montet, Christian and Serra, Daniel. 2003. *Game theory and economics.* Basingstoke, Hampshire [u.a.]: Palgrave MacMillan.

Morgan, Jerry L. 1978. "Two types of convention in indirect speech acts." In Peter Cole (ed.), *Pragmatics*, volume 9 of *Syntax and semantics*, 261–280. New York: Academic Press.

Nash, John Forbes. 1952. “Non-cooperative games.” *The Annals of Mathematics* 52:286–295.

Neale, Stephen R. A. 1992. “Paul Grice and the philosophy of language.” *Linguistics and Philosophy* 15:509–559.

Neumann, John von and Morgenstern, Oskar. 1953. *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press, 3 edition.

Niedzielski, Nancy A. and Preston, Dennis Richard. 2000. *Folk linguistics*. Berlin: Mouton de Gruyter.

Nimtz, Christian. 2009. “Conceptual truth defended.” In Nikola Kompa, Christian Nimtz, and Christian Suhm (eds.), *The a priori and its role in philosophy*, 137–155. Paderborn: Mentis.

Nimtz, Christian and Beckermann, Ansgar (eds.). 2005. *Philosophie und/als Wissenschaft: Hauptvorträge und Kolloquiumsbeiträge zu GAP.5 ; Fünfter Internationaler Kongress der Gesellschaft für Analytische Philosophie, Bielefeld, 22. - 26. September 2003 = Philosophy - Science - Scientific Philosophy*. Perspektiven der Analytischen Philosophie. Paderborn: Mentis.

Nolan, Daniel. 2005. *David Lewis*. Philosophy now. Chesham: Acumen.

Norman, Guy. 2002. “Description and prescription in dictionaries of scientific terms.” *International Journal of Lexicography* 15:259–276.

Nowak, Martin A. 2000. “Evolutionary biology of language.” *Philosophical Transactions of the Royal Society B: Biological sciences* 355:1615–1622.

Nowak, Martin A. and Krakauer, David C. 1999. “The evolution of language.” *Proceedings of the National Academy of Sciences of the United States of America* 96:8028–8033.

Nowak, Martin A., Krakauer, David C., and Dress, Andreas. 1999a. “An error limit for the evolution of language.” *Proceedings of the Royal Society B: Biological Sciences* 266:2131–2136.

Nowak, Martin A., Plotkin, Joshua B., and Jansen, Vincent A. A. 2000. “The evolution of syntactic communication.” *Nature* 404:495–498.

Nowak, Martin A., Plotkin, Joshua B., and Krakauer, David C. 1999b. “The evolutionary language game.” *Journal of Theoretical Biology* 200:147–162.

O’Leary-Hawthorne, John. 1990. “A note on ’Languages and language’.” *Australasian Journal of Philosophy* 68:116–118.

—. 1993. “Meaning and evidence: A reply to Lewis.” *Australasian Journal of Philosophy* 71:206–211.

Pagin, Peter and Pelletier, Francis J. 2007. “Content, context, and composition.” In Gerhard Preyer and Georg Peter (eds.), *Context-sensitivity and semantic minimalism.* Oxford: Oxford University Press.

Peacocke, Christopher. 1976. “Truth definitions and actual languages.” In Evans and McDowell (1976), 162–188.

Pink, Thomas. 2004. “Moral obligation.” In Anthony O’Hear (ed.), *Modern moral philosophy*, volume 54 of *Royal Institute of Philosophy supplement*, 159–188. Cambridge: Cambridge University Press.

Plato. 1969. *Dialogues of Plato: Translated into English with analyses and introductions*, volume 3. Oxford: Clarendon Press, 4 edition.

Platts, Mark Bretton de. 1997. *Ways of meaning: An introduction to a philosophy of language.* A Bradford book. Cambridge, MA: MIT Press, 2 edition.

Posner, Eric A. 2000. *Law and social norms.* Cambridge, MA [u.a.]: Harvard University Press.

Putnam, Hilary. 1981. *Reason, truth and history.* Cambridge: Cambridge University Press.

—. 1983. “Models and reality.” In Hilary Putnam (ed.), *Philosophical papers*, 1–25. Cambridge: Cambridge University Press.

—. 1997. “The meaning of ”meaning”.” In Hilary Putnam (ed.), *Mind, language and reality*, volume Vol. 2 of *Philosophical papers / Hilary Putnam*, 215–271. Cambridge: Cambridge University Press, 2 edition.

Quine, Willard V. O. 1960. *Word and object.* Studies in communication. Cambridge, MA: MIT Press.

Quine, Willard V. O. (ed.). 1976a. *The ways of paradox and other essays.* Cambridge, MA: Harvard University Press, 2 edition.

Quine, Willard V. O. 1976b. “Truth by convention.” In Quine (1976a), 77–106.

—. 1980. “Two dogmas of empiricism.” In Willard V. O. Quine (ed.), *From a logical point of view*, 20–46. Cambridge: Harvard University Press, 2 edition.

Recanati, François. 2004. *Literal meaning.* Cambridge: Cambridge University Press.

Reimer, Marga. 2004. "What malapropisms mean: a reply to Davidson." *Erkenntnis* 60:317–334.

Rescorla, Michael. 2007. "Convention." In Zalta (2007).

Resnik, Michael D. 2002. *Choices: An introduction to decision theory.* Minneapolis: University of Minnesota Press.

Richerson, Peter J. and Boyd, Robert. 2006. *Not by genes alone: How culture transformed human evolution.* Chicago [u.a.]: University of Chicago Press.

Rolls, Edmund T. 2004. *The brain and emotion.* Oxford: Oxford University Press, 1 edition.

Russell, Bertrand. 1921. *The analysis of mind.* London [u.a.]: Unwin Brothers Ltd.

Savage, Leonard J. 1972. *The foundations of statistics.* Toronto: Dover Publications, 2 edition.

Savigny, Eike von. 1983. *Zum Begriff der Sprache: Konvention, Bedeutung, Zeichen.* Stuttgart: Reclam.

—. 1985. "Social habits and enlightened cooperation: Do humans measure up to Lewis conventions?" *Erkenntnis* 22:79–96.

—. 1988. *The social foundations of meaning.* Berlin [u.a.]: Springer.

Schank, Roger C. and Abelson, Robert P. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* The artificial intelligence series. New York: Wiley.

Schelling, Thomas C. 2005. *The strategy of conflict.* Cambridge, MA: Harvard University Press.

Schiffer, Stephen R. 1972. *Meaning.* Oxford: Clarendon Press.

—. 1993. "Actual-language relations." *Philosophical Perspectives* 7:231–258.

—. 2006. "Two perspectives on knowledge of language." *Philosophical Issues* 16:275–287.

Schulte, Joachim. 1993. "The uses of mistakes." In Stoecker (1993), 149–157.

Schulte, Peter. 2008. *Zwecke und Mittel: Eine expressivistische Theorie der instrumentellen Rationalität.* Ph.D. thesis, Universität Bielefeld, Bielefeld.

—. 2009. “Moral and rational ”oughts”: the distinction between ”demanding” and ”recommending” normativity.” In Dora Achourioti, Edgar Andrade, and Marc Staudacher (eds.), *Proceedings for the graduate philosophy conference on normativity*, number X-2009-02 in ILLC Publications, 161–168. Amsterdam.

—. 2010. “Truthmakers: a tale of two explanatory projects.” *Synthese* .

Schwarz, Wolfgang. 2009. *David Lewis: Metaphysik und Analyse.* Paderborn: Mentis.

Searle, John R. 1969. *Speech acts: an essay in the philosophy of language.* Cambridge: Cambridge University Press.

Searle, John R. and Vanderveken, Daniel. 1985. *Foundations of illocutionary logic.* Cambridge [u.a.]: Cambridge University Press.

Shea, Nicholas. forthcoming. “Millikan’s isomorphism requirement.” In Justine Kingsbury, Dan Ryder, and Ken Williford (eds.), *Millikan and her critics.* Oxford: Blackwell.

Sheridan, Richard Brinsley. 1775. *The rivals, a comedy: As it is performed at the Theatre-Royal in Covent-Garden.* Dublin.

Sick, Bastian. 2004. “zeitgleich/gleichzeitig.”

Sillari, Giacomo. 2008. “Common knowledge and convention.” *Topoi* 27:29–39.

Skyrms, Brian. 1994. “Darwin meets the Logic of Decision: Correlation in Evolutionary Game Theory.” *Philosophy of Science* 61:503–528.

—. 1998. *Evolution of the social contract.* Cambridge: Cambridge University Press.

—. 2010. *Signals: Evolution, learning, & information.* Oxford [u.a.]: Oxford University Press.

Sobel, Joel. 2009. “Signaling games.” In Robert A. Meyers (ed.), *Encyclopedia of Complexity and Systems Science.* New York: Springer, 1 edition.

Speaks, Jeffrey. 2010. “Theories of meaning.” In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition).*

Spence, Michael. 1973. “Job market signaling.” *The Quarterly Journal of Economics* 87:355–374.

Sperber, Dan and Wilson, Deirdre. 1986. “Loose talk.” *Proceedings of the Aristotelian Society* 86:153–171.

—. 1995. *Relevance: Communication and cognition.* Oxford, Cambridge: Blackwell, 2 edition.

—. 2002. “Truthfulness and relevance.” *Mind* 111:583–632.

Sripada, Chandra and Stich, Stephen P. 2006. “A framework for the psychology of norms.” In Peter Carruthers (ed.), *The innate mind*, Evolution and cognition, 280–301. New York, NY: Oxford University Press.

Stainton, Robert J. 1993. *Non-sentential assertions.* Ph.D. thesis, Massachusetts Institute of Technology.

—. 1994. “Using non-sentences: an application of Relevance Theory.” *Pragmatics & Cognition* 2:269–284.

—. 2004. “In defense of non-sentential assertion.” In Zoltán G. Szabó (ed.), *Semantics vs. Pragmatics.* Oxford: Clarendon Press.

—. 2006. *Words and thoughts: Subsentences, ellipsis, and the philosophy of language.* Oxford: Clarendon Press.

Staudacher, Marc. 2007. *Ein sprachphilosophisches Problem mit dialogeinleitenden Fragmenten.* Ph.D. thesis, Universität Bielefeld, Bielefeld.

Stemmer, Peter. 2008. *Normativität: Eine ontologische Untersuchung.* Berlin: Walter de Gruyter.

Stenius, Erik. 1967. “Mood and language game.” *Synthese* 17:254–274.

Stoecker, Ralf (ed.). 1993. *Reflecting Davidson: Donald Davidson responding to an international forum of philosophers.* Grundlagen der Kommunikation und Kognition. Berlin: Walter de Gruyter.

Stoljar, Daniel. 2009. “Physicalism.” In Zalta (2009).

Sugden, Robert. 1986. *The economics of rights, co-operation and welfare.* Oxford: Blackwell.

—. 1998. “The role of inductive reasoning in the evolution of conventions.” *Law and Philosophy* 17:377–410.

—. 2004. *The economics of rights, co-operation and welfare.* Basingstoke: Palgrave MacMillan, 2 edition.

Syverson, Paul F. 2003. *Logic, convention, and common knowledge: A conventionalist account of logic*, volume 142 of *CSLI lectures notes.* Stanford, CA: CSLI Publications.

Szabó, Zoltán G. 2008. "Structure and conventions." *Philosophical Studies* 137:399–408.

The Royal Swedish Academy of Sciences. 2005. "Robert Aumann's and Thomas Schelling's contributions to game theory: Analyses of conflict and cooperation."

Tielmann, Christian. 2005. *Sprachregeln und Idiolekte: Plädoyer für einen normativistischen Individualismus.* Ph.D. thesis, Universität Hamburg, Hamburg.

Trapa, Peter E. and Nowak, Martin A. 2000. "Nash equilibria for an evolutionary language game." *Journal of Mathematical Biology* 41:172–188.

Trivers, Robert L. 1971. "The evolution of reciprocal altriusm." *The Quarterly review of biology* 46:35–57.

Tsur, Oren, Davidov, Dmitry, and Rappoport, Ari. 2010. "ICWSM – A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 162–169. Menlo Park, CA: AAAI Press.

Ullmann-Margalit, Edna. 1977. *The emergence of norms.* Clarendon library of logic and philosophy. Oxford: Clarendon Press.

van Rooy, Robert. 2003. "Quality and quantity of information exchange." *Journal of Logic, Language and Information* 12:423–451.

Vanderschraaf, Peter and Sillari, Giacomo. 2007. "Common knowledge." In Zalta (2007).

Weatherson, Brian. 2009. "David Lewis." In Zalta (2009).

Weber, Max. 1997. *The theory of social and economic organization: Translated to English by Talcott Parsons.* New York, NY: Free Press, 1 edition.

Whiting, Daniel. 2007. "The normativity of meaning defended." *Analysis* 67:133–140.

—. 2009. "Is meaning fraught with." *Pacific Philosophical Quarterly* 90:535–555.

Wikforss, Åsa M. 2001. "Semantic normativity." *Philosophical Studies* 102:203–226.

Wikipedia. 09.06.2010. "Heap spraying."

—. 10.03.2009. "Paul Revere."

—. 11.03.2009. "Chicken (game)."

—. 12.05.2010. “Nix v. Hedden.”

—. 13.03.2009. “Old North Church.”

—. 19.02.2010. “Homonym.”

—. 28.01.2009. “Common knowledge (logic).”

Williams, Bernard A. O. 2002. *Truth & truthfulness: An essay in genealogy.* Princeton, N.J. [u. a.]: Princeton University Press.

Wimmer, Heinz and Perner, Josef. 1983. “Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception.” *Cognition* 13:103–128.

Winch, Peter. 1963. *The idea of a social science and its relation to philosophy.* London [u.a.]: Routledge & Kegan Paul.

Wittgenstein, Ludwig and Waismann, Friedrich. 2003. *The Voices of Wittgenstein: The Vienna Circle: original German texts and English translations.* London: Routledge.

Young, H. Peyton. 1998. *Individual strategy and social structure: An evolutionary theory of institutions.* Princeton, NJ: Princeton University Press.

Zalta, Edward N. (ed.). 2007. *The Stanford Encyclopedia of Philosophy (Fall 2007 Edition).*

—. 2009. *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition).*

Zuidema, Willem H. 2005. *The major transitions in the evolution of language.* Ph.D. thesis, University of Edinburgh, Edinburgh.

Index of key terms

Appendix A

List of named theses

In this appendix, a list of named theses is provided for reference purposes. For each thesis, all occurrences are listed with a page reference.

A1. If x believes y, then x recognizes y. *Occurrences: p. 150.*

A1′. If x expects y, then x recognizes y. *Occurrences: p. 150.*

A2. x expects y iff x has reason to believe y and x is rational to the (high enough) degree n. *Occurrences: p. 150.*

A3. I reason rationally to a high enough degree. *Occurrences: p. 150.*

A4. I expect that you expect that I am rational. *Occurrences: p. 151.*

A5. I expect that you are rational. *Occurrences: p. 151.*

ALR1. The meaning of an expression in a natural language of a population is determined by the population's use of an abstract language corresponding to the natural language. *Occurrences: p. 157.*

ALR2. A population uses an abstract language $\mathcal{L}$ iff there are conventions of truthfulness and trust in $\mathcal{L}$ among the members of the population. *Occurrences: p. 158.*

ALR3. There are conventions of truthfulness and trust in an abstract language $\mathcal{L}$ among members of a population P iff speakers of P try to avoid uttering sentences not true in $\mathcal{L}$ (truthfulness) and hearers of P tend to believe that sentences uttered by speakers of P are true in $\mathcal{L}$ (trust). *Occurrences: p. 158.*

ALR-E. Expression e means m in $\mathcal{L}_\Gamma$ of P iff P uses Γ and Γ pairs e with m. *Occurrences: p. 162.*

ALR-E′. Expression e means m in $\mathcal{L}_\mathcal{T}$ of P iff (i) P uses $\mathcal{L}$, (ii) $\mathcal{L} \subseteq \mathcal{L}_\mathcal{T}$, (iii) members of P process $\mathcal{L}$-utterances via an $\mathcal{L}$-determining translator $\mathcal{T}$, (iv) $\mathcal{T}$ generates $\mathcal{L}_\mathcal{T}$, and (v) $\mathcal{T}$ pairs e with m. *Occurrences: p. 170.*

ALR-S. Sentence s means m in $\mathcal{L}$ used by P iff P uses $\mathcal{L}$ and $\mathcal{L}(s) = m$. *Occurrences: p. 162.*

C. For all expressions e, meanings m, coordinates $\mathcal{C}$: The conventional use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$. *Occurrences: p. 6, 29, 32.*

C′. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$. *Occurrences: p. 32, 171, 194.*

C″. For all expressions e, meanings m, coordinates $\mathcal{C}$: The stable linguistic use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$. *Occurrences: p. 172, 194, 293.*

C$_-$N$_-$. Meaning is neither conventional nor normative. *Occurrences: p. 6.*

C$_-$N$_+$. Meaning is not conventional but normative. *Occurrences: p. 6.*

C$_+$N$_-$. Meaning is conventional but not normative. *Occurrences: p. 6.*

C$_+$N$_+$. Meaning is both conventional and normative. *Occurrences: p. 6.*

C0. A convention is *social*: there is a group G of agents. *Occurrences: p. 22.*

C1. A convention *involves a pattern of activity*: (i) there is a pattern R of individual activities of members of G, (ii) R determines for a range of activities whether they are conforming or deviating, and (iii) at least on some occasions, members of G would behave in a way conforming to R. *Occurrences: p. 22.*

C2. A convention *requires coordination*: Members of G have (together) an effective coordinative disposition to behave in a way conforming to R. *Occurrences: p. 22.*

C3. A convention is *relatively robust*: in all near futures, it exists as well. *Occurrences: p. 22.*

C-M. If there is a conventional SCH-pattern for description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among members of g at t in w. *Occurrences: p. 276.*

C-NC. Other things being equal, when a convention becomes entrenched among human members, it becomes a normative convention. *Occurrences: p. 248.*

CS. There is a normal situation in which a sender of a conventional signaling system produces ("utters") a signal in a way conforming to the signaling convention. *Occurrences: p. 147.*

CS′. There is a normal situation in which a sender of a conventional signaling system produces ("utters") a signal in a way conforming to the signaling convention iff

a. There is a regularity R in situations which are two-sided signaling problems G among you and me, and R is a conventional signaling system $\langle \sigma, \rho \rangle$ among us.
b. Let m be a message, t a state the world can be in, and a a response such that $m = \sigma(t)$ and $a = \rho(m)$.
c. There is a situation t which is a two-sided signaling problem of type G where I am the sender and you are the receiver.
d. I have observed that t holds.
e. I signal m in conformity to our convention.
f. The convention is "perfect," that is, there are no exceptions to it. *Occurrences: p. 149.*

C-SLU. If there is a conventional SCH-pattern for description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w, then there is a stable linguistic use of δ indicating μ with derivational complexity z among members of g at t in w. *Occurrences: p. 277.*

DSN1. The enforcers have power over the addressees at t in w. *Occurrences: p. 246.*

DSN2. The enforcers tend to sanction the addressees' R-concerning behavior at t in w. *Occurrences: p. 246.*

DSN3. The addressees tend to behave according to R at least partly because of (DSN2) the enforcers' tendency to sanction their R-concerning behavior at t in w. *Occurrences: p. 246.*

FM1. *First meaning is systematic.* A competent speaker or interpreter is able to interpret utterances, his own or those of others on the basis of the semantic properties of the parts, or words, in the utterance, and the structure of the utterance. For this to be possible, there must be systematic relations between the meanings of utterances. *Occurrences: p. 77.*

FM2. *First meanings are shared.* For speaker and interpreter to communicate successfully and regularly, they must share a method of interpretation of the sort described in FM1. *Occurrences: p. 77.*

FM3. *First meanings are governed by learned conventions or regularities.* The systematic knowledge or competence of the speaker or interpreter is learned in advance of occasions of interpretation and is conventional in character. *Occurrences: p. 77.*

G1. We express normative attitudes with normative statements. *Occurrences: p. 237.*

G2. Normative attitudes are a special type of conative mental states. *Occurrences: p. 237.*

G3. Since normative attitudes are conative, normative statements do not have a descriptive but an "imperative" character. *Occurrences: p. 237.*

GS. A-facts supervene globally on B-facts iff any world which is a minimal B-duplicate of our world is also an A-duplicate of our world. *Occurrences: p. 13.*

H. For any language $\mathcal{L}$: The meaning of (almost) any expression of $\mathcal{L}$ depends on the meanings of (almost) all other expressions of $\mathcal{L}$; *i.e.* each difference between $\mathcal{L}$ and another language $\mathcal{L}'$ implies that the expressions in $\mathcal{L}$ have other meanings than the expressions in $\mathcal{L}'$. *Occurrences: p. 207.*

HD. If it's the use of expressions in a community which determines their meanings, then in which way does this determination depend on the members and the circumstances? *Occurrences: p. 36, 288.*

I1. According to conventionalist accounts, the literal meaning of an expression on a particular occasion of use is its *specific first meaning. Occurrences: p. 78.*

I2. A necessary condition for successful linguistic communication is that the hearer understands the expression the speaker uttered in its generic first meaning. *Occurrences: p. 78.*

I3. There are cases (occasions) of successful linguistic communication with malapropisms. *Occurrences: p. 80.*

I4. According to conventionalist accounts, in case of successful linguistic communication with a malapropism, a hearer understands the malapropism the speaker uttered in its specific first meaning. *Occurrences: p. 80.*

I5. In cases of successful communication with malapropisms, the malapropism's generic first meaning is not identical to its contextually-relativized conventional meaning. *Occurrences: p. 81.*

I6. In cases of successful communication with malapropisms, if a hearer understands an expression a speaker uttered in its generic first meaning, then she doesn't understand it in its (contextually-relativized) conventional meaning (and *vice versa*). *Occurrences: p. 81.*

JSN3. Each enforcer could accept a system of norms of which N is a part at t in w. *Occurrences: p. 245.*

LSM1. I signal m with the intention that you do a. *Occurrences: p. 148.*

LSM2. I expect you to recognize my intention that you do a, when you observe that I signal m. *Occurrences: p. 148.*

LSM2′. I expect that (you expect that (I intend that you do a), when you observe that I signal m). *Occurrences: p. 150.*

LSM2′a. I expect that (you expect that t holds, when you observe that I signal m). *Occurrences: p. 151.*

LSM2′b. I expect that you expect that I desire that you do a, conditionally upon t. *Occurrences: p. 151.*

LSM2′c. I expect that you expect that (I desire that you do a, when you observe that I signal m). *Occurrences: p. 151.*

LSM2′d. I expect that you expect that I expect that (you do a, when you observe that I signal m). *Occurrences: p. 151.*

LSM3. I expect your recognition of my intention to be effective in bringing it about that you do a. I do not regard it as a foregone conclusion that my action will bring it about that you do a, whether or not you recognize my intention that you do a. *Occurrences: p. 148.*

LSN1. There is a stable linguistic use among the arbitrators at time t in world w whose pattern of activity is Θ. *Occurrences: p. 278.*

LSN2. There is a system of norms N according to which the addressees are N-required to conform to Θ and N-forbidden to deviate from Θ. *Occurrences: p. 279.*

LSN3. There is a rationalistic social norm to conform to Θ among the addressees enforced by the enforcers accepting N at time t in world w. *Occurrences: p. 279.*

MD-ME. If meaning is normative, then this is because of the way it is constituted (determined). *Occurrences: p. 41.*

MI1. Expression types are reproductively established families of expression tokens. That is, expressions are individuated by the reproduction chain of patterns in which the expressions occur. *Occurrences: p. 198.*

MI2. Meaningful expressions don't have a meaning relative to a language. *Occurrences: p. 199.*

MI3. Expression e means semantic mapping function $\mathcal{M}$ iff (i) e has the stabilizing function ϕ and (ii) for all tokenings t of e: if t performs ϕ, then t maps onto world affairs according to $\mathcal{M}$. *Occurrences: p. 202.*

MI4. Expression e means semantic mapping function $\mathcal{M}$ iff e's use pattern is a coordination convention which proliferates because (i) e has the stabilizing function ϕ and (ii) for all tokenings t of e: if t performs ϕ, then t maps onto world affairs according to $\mathcal{M}$. *Occurrences: p. 205.*

N. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$, then utterances of "e means m at $\mathcal{C}$" can be used to express an *ought* with a demanding character. *Occurrences: p. 6, 17, 40, 65.*

N′. For all expressions e, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use e in accordance with its meaning m at $\mathcal{C}$ (by uttering "e means m at $\mathcal{C}$" which expresses an *ought* with a demanding character). *Occurrences: p. 65, 284, 293.*

N″. For all descriptions δ, meanings m and coordinates $\mathcal{C}$: If e means m at $\mathcal{C}$ in virtue of a social norm, then the enforcers of the social norm can demand that the addressees use δ in accordance with its meaning m at $\mathcal{C}$ (by uttering "δ means m at $\mathcal{C}$" which expresses an *ought* with a demanding character). *Occurrences: p. 284.*

N1. It's a practical *must*. The object of the *must* are actions. Sometimes, the object can also be a state, the possession of a property, or the attaining of something, on condition that it is one's own acting that gets oneself in such a state or in the possession of the properties. *Occurrences: p. 14.*

N2. A normative *must* does not rule out that one acts differently than how one must. A normative *must* is not a force that inevitably moves (or will or would move) a person all the way to action. *Occurrences: p. 14.*

N3. Normative *musts* are tied to a pressure to act. It presses its addressee to do certain actions. *Occurrences: p. 15.*

N4. The normative *must* is always ontologically subjective. Its existence depends on thinking, feeling, and wanting of humans (or other living creatures). *Occurrences: p. 15.*

NA1. Normative attitudes are logically combinable. Example: Suppose I N-want to clean the kitchen on Sunday. I also N-want to write a letter on Tuesday. Then I N-want to do both. *Occurrences: p. 238.*

NA2. Normative attitudes are under consistency pressure. Example Suppose I N-want to clean the kitchen on Sunday and to go to the football match. But only one goal can be realized. Hence I N-want to give up one of my N-wants. *Occurrences: p. 239.*

NA3. Normative attitudes are generalizable. Example: Suppose *I* N-want you not to harm other people. Then I N-want *everyone* not to do harm to them. *Occurrences: p. 239.*

NA4. Normative attitudes are embedded in a hierarchy of higher-order normative attitudes. Example: Suppose I N-want to clean the kitchen on Sunday. Then I N-want to have this N-want. *Occurrences: p. 239.*

NDP. An agent in a decision problem under uncertainty ought to perform a subjectively rational strategy. *Occurrences: p. 97.*

NE. A strategy profile s is a Nash equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) \geq u_2(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) \geq u_1(s'')$. *Occurrences: p. 98.*

PCE. A strategy profile s is a proper coordination equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) > u_2(s')$ $\underline{\text{and}}$ $u_1(s) > u_1(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) > u_1(s'')$ $\underline{\text{and}}$ $u_2(s) > u_2(s'')$. *Occurrences: p. 99.*

PF. A thing x has the proper function to ϕ iff x exists and is the way it is because the ancestors of x have performed ϕ sufficiently often. *Occurrences: p. 197.*

RSN1. The enforcers have power over the addressees at t in w. *Occurrences: p. 232, 235.*

RSN1′. For all enforcers e and all addressees g: e has power over g at t in w. *Occurrences: p. 235.*

RSN1″. The enforcers could form a coalition that would together have power over any coalition the addressees could possibly form. *Occurrences: p. 235.*

RSN2. According to N, the addressees are N-required to conform to R and N-forbidden to deviate from R. *Occurrences: p. 232, 239.*

RSN3. Each enforcer accepts a system of norms of which N is a part at t in w. *Occurrences: p. 232, 240.*

RSN4. The enforcers tend to sanction the addressees' R-concerning behavior at least partly because of (RSN3) their accepting N at t in w. *Occurrences: p. 232, 240.*

RSN5. The addressees tend to behave according to R at least partly because of (RSN4) the enforcers' tendency to sanction their R-concerning behavior at t in w. *Occurrences: p. 232, 240.*

S0. A social norm is *social*: there are various not necessarily disjoint groups including (i) a group E of enforcers and (ii) a group G of addressees. *Occurrences: p. 26, 231.*

S1. A social norm *involves a pattern of activity*: (i) there is a pattern R of individual activities of the addressees, (ii) R determines for a range of activities whether they are conforming or deviating, and (iii) at least on some occasions, the addressees would behave in a way conforming to R. *Occurrences: p. 26, 231.*

S2. A social norm is *prescriptive*: A social norm is, in part, constituted by a norm N which determines for a range of activities whether they are prescribed, forbidden, or allowed. The norm N prescribes to conform to R. N is enforced by the enforcers who accept it and have power over the addressees. *Occurrences: p. 26, 231.*

S3. A social norm has a *demanding character*: The enforcers are in a position to demand conformity to R from the addressees. *Occurrences: p. 27, 231.*

S4. A social norm is *relatively robust*: in all near futures, the enforcers accept the same norm and at least on some occasions, the addressees behave in a way conforming to R. *Occurrences: p. 27, 232.*

SG1. A theoretically interesting part of natural languages can be explained by signaling systems. *Occurrences: p. 124.*

SG2. Important parts of a signaling system are: (i) signals, (ii) a population, (iii) a sender and receiver role, (iv) states of affairs observable by senders, (v) reactions of receivers, (vi) a pattern of activity prevailing in the population to use *signals* according to certain *contingency plans.* The contingency plans for the members of the population consist of two parts, one for the sender-role and one for the receiver-role.

a. A sender's plan determines which signal to produce depending on what the sender has observed, her desires, and her beliefs.

b. A receiver's plan determines how to react to an observed signal depending on her desires and beliefs. *Occurrences: p. 124.*

SG3. Signals in a signaling system have a meaning in virtue of the fact that there is a pattern of activity that is a convention in the population. Which meanings the signals have depends on the pattern which is the convention. *Occurrences: p. 124.*

SG-G. Normally, senders speaker-mean something (in *Grice*'s sense) when they utter a verbal expression conforming to a signaling convention they are party to. *Occurrences: p. 147.*

SLU-M. If there is a stable linguistic use of description δ indicating content+ μ with derivational complexity z among members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among members of g at t in w. *Occurrences: p. 277.*

SNE. A strategy profile s is a strict Nash equilibrium iff for all other strategy profiles s' which agree on what player 1 does in s, $u_2(s) > u_2(s')$, and for all other strategy profiles s'' which agree on what player 2 does in s, $u_1(s) > u_1(s'')$. *Occurrences: p. 98.*

SN-M. If there is a linguistic social norm for description δ indicating content+ μ with derivational complexity z addressed to members of group g at time t in world w and δ's derivational complexity does not exceed z, then δ means μ among the members of g at t in w. *Occurrences: p. 280.*

SN-SLU. If there is a linguistic social norm for description δ indicating content+ μ with derivational complexity z addressed to members of group g at time t in world w, then there is a stable linguistic use of δ indicating μ with derivational complexity z among members of g at t in w. *Occurrences: p. 280.*

SP. Sub-sentential expressions have the meanings they do *because* sentences in which they occur mean what they do. *Occurrences: p. 185.*

SRB. An action a in a decision problem under uncertainty is *subjectively rational* for its agent iff a is among the actions with the maximal expected utility, that is, $a \in \{a' | \neg\exists a'' : EU(a'') > EU(a')\}$. *Occurrences: p. 97.*

U. For all expressions e, meanings m, coordinates $\mathcal{C}$: The use of e at $\mathcal{C}$ determines that e means m at $\mathcal{C}$. *Occurrences: p. 5.*

VA. Conventions are verbally performed agreements. *Occurrences: p. 145.*

Appendix B

List of adequacy conditions

The adequacy conditions were introduced in §1.4. For reference purposes, they are listed here again.

Conventions

DesC1. The account must be faithful to the pre-theoretic characterization of a convention (C0–C3 in §1.2). *Occurrences: p. 33.*

DesC2. The account must provide an answer to the question what conventions are (*e.g.* behavioral regularities, rules, patterns). *Occurrences: p. 33.*

DesC3. The account must provide a taxonomy of the kinds of conventions there are. *Occurrences: p. 33.*

DesC4. The account must provide an answer to the question whether (and if so which) epistemic states are involved among the parties to a convention. *Occurrences: p. 33.*

DesC5. It must be possible that in a human population conventions are created, learned, sustained, and changed. *Occurrences: p. 33.*

DesC6. The dynamics of conventions must be explained. *Occurrences: p. 33.*

Social norms

DesN1. The account must be faithful to the pre-theoretic characterizations of a social norm (S0–4 in §1.2) and of normativity (N1–3 in §1.1.3). *Occurrences: p. 34, 253.*

DesN2. The account must provide an answer to the question what (social) norms and normativity are. *Occurrences: p. 34, 253.*

DesN3. The account must provide a taxonomy of the kinds of (social) norms there are. *Occurrences: p. 34, 253.*

DesN4. The account must provide an answer to the question what kind of epistemic states are involved in a (social) norm. *Occurrences: p. 34, 254.*

DesN5. It must be possible that in a human population social norms are created, learned, sustained, and changed. *Occurrences: p. 34, 254.*

DesN6. The dynamics of (social) norms must be explained. *Occurrences: p. 34, 254.*

Conventionalist accounts of meaning

DesM1. The account must be coherent with an explanation of the communicative functions of meaning-sentences. *Occurrences: p. 34, 291.*

DesM2. The account must explain the meanings of expressions in terms of their stable uses (conventions, social norms, and normative conventions). *Occurrences: p. 34, 291.*

DesM3. The account must provide a plausible notion of a semantic mistake. *Occurrences: p. 34, 291.*

DesM4. The account must allow for a plausible conception of a public language. *Occurrences: p. 35, 291.*

DesM5. The account must explain the usual meaning facts. *Occurrences: p. 35, 291.*

DesM6. The account must provide a solution to Humpty Dumpty's problem. *Occurrences: p. 35, 292.*

Samenvatting

Dit proefschrift is een bijdrage aan de taalfilosofie. De centrale vraag is: op grond van welke feiten betekenen talige expressies wat ze betekenen? Waarom bijvoorbeeld betekent "appel" *appel* in het Nederlands? De vraag krijgt een systematisch antwoord, te weten: talige expressies betekenen wat ze betekenen omdat er, onder hun gebruikers, talige conventies en sociale normen bestaan, die expressies op een bepaalde manier gebruiken en begrijpen.

Het antwoord wordt verklaard en verdedigd als een centrale stelling. In deze vorm is het op zijn best een slogan: Wat is betekenis? Wat is het, expressies gebruiken en begrijpen? Wat zijn conventies en sociale normen eigenlijk? Hoe bepaalt het gebruik en het begrip betekenis? Het doel van het proefschrift is het beantwoorden van deze vragen.

In hoofdstuk 1 wordt het project waarin deze vragen aan bod komen, uitgelegd en gemotiveerd. De theorieën worden verdeeld in drie basistypes (of paradigma's). Er wordt onderscheid gemaakt tussen sociale normen en conventies en er wordt een voorwaarde gesteld voor een adequate theorie. Op basis van deze voorwaarde kunnen we dit soort theorieën evalueren.

In hoofdstuk 2 wordt een volgende stelling onderzocht die van belang is voor de adequaatheid van dit soort theorieën: de zogenoemde *stelling van de "normativiteit van betekenis"*. Als een expressie iets betekent, dan zou men deze expressie op een bepaalde manier gebruiken en begrijpen. De stelling wordt onder voorbehoud geaccepteerd.

In hoofdstuk 3 wordt het project verdedigd tegen een fundamentele tegenwerping van Donald Davidson. Volgens hem zijn conventies op een bepaalde manier niet essentieel voor betekenis.

In de hoofdstukken 4 tot en met 9 worden de theorieën van de drie paradigma's kritisch besproken:

- Signaal spelen: Op dit moment zijn theorieën van het eerste paradigma onderwerp van actief onderzoek in de speltheorie. Volgens deze zijn taalgebruikers actoren die, of als sprekers observeren en dan een signaal geven, of als luisteraars de signalen observeren waarna ze op typische manieren reageren. Er zijn twee standaard interpretaties van dit soort theorieën. Volgens de rationele interpretatie overwegen actoren welk signaal te geven en hoe te reageren op een observatie. Volgens de alternatieve interpretatie hebben de actoren de dispositie om te signaleren en daarop te reageren op dezelfde manier als waarop de dispositie wordt gerealiseerd (ze zijn kenmerkend voor het resultaat van het leren).
- Actuele taal relaties: Theorieën van het tweede paradigma gelden als standaard in de analytische filosofie. Als bouwstenen gebruiken zij taaltheorieën van het soort dat taalkundigen ontwikkelen. Zo'n soort bouwsteen is gerelateerd aan de sociale praktijken van het taalgebruik in een gemeenschap.
- Evolutionaire theorieën: Theorieën van het derde paradigm vatten taal op als een cluster van conventionele gedragingen die evolutionair in het voordeel zijn (in de brede zin van het woord "evolutie", inclusief cultuur en biologie). Een belangrijke vertegenwoordiger van zo'n theorie is Ruth Millikan.

In de hoofdstukken 4 tot en met 6 worden de recente theorieën van de eerste twee paradigma's besproken (meestal in een rationele interpretatie); deze gaan terug op het werk van David Lewis. Er worden verschillende problemen geconstateerd die moeilijk zijn op te lossen. Ik pleit voor een theorie van het derde paradigma. Deze theorieën worden in de hoofdstukken 7 tot en met 9 besproken.

In hoofdstuk 7 worden de theorieën van Ruth Millikan en Simon Huttegger geëvalueerd, waarbij de laatste een evolutionaire theorie van signaalspellen is. Daarbij worden enige problemen en beperkingen geconstateerd. Een van de problemen is dat een adequate theorie over sociale normen ontbreekt.

In hoofdstuk 8 wordt een theorie over sociale normen ontwikkeld. Zo'n theorie is nodig om de stelling "normativiteit van betekenis" te onderbouwen. Op basis van deze theorie wordt, alhoewel niet volledig, één van de centrale vragen van dit proefschrift beantwoord.

In hoofdstuk 9 wordt een alternatieve conventionalistische theorie ontwikkeld. Deze theorie maakt gebruik van de theorie van conventies van Millikan, van mijn theorie over sociale normen, en van een nieuwe bepaling van communicatie die het mogelijk maakt woorden direct betekenissen toe te kennen. Mijn alternatieve theorie realiseert een unieke combinatie van positieve kenmerken die hem daardoor relatief gezien beter maken dan de alternatieve conventionalistische theorieën die in dit proefschrift besproken worden.

In hoofdstuk 10 worden de belangrijkste stellingen van het proefschrift samengevat en worden vragen voor toekomstig onderzoek opgesomd.

Abstract

This dissertation is a contribution to the philosophy of language. Its central question is: *In virtue of which facts do linguistic expressions mean what they do? E.g.* why does "apple" mean *apple* in English? The question receives a systematic answer; in short: Linguistic expressions mean what they do because among their users, there are linguistic conventions and social norms to use and understand them in certain ways.

The answer is clarified and defended as a central thesis. For in this form, it is at best a slogan: What is meaning? What is it to use and understand expressions? What are conventions and social norms anyway? How does the use and understanding determine meaning? The goal of the dissertation consists in answering these questions.

In chapter 1, the project these questions belong to is explained and motivated. Three basic types (or paradigms) of accounts are distinguished, a distinction between conventions and social norms is introduced, and an adequacy condition is proposed. Thereby we're in a position to evaluate such theories.

In chapter 2, a further thesis is examined which is important for the adequacy of such theories: The so-called "normativity of meaning" thesis: If an expression means something, then there is an *ought* concerning its use and understanding. With an important restriction, the thesis is accepted.

In chapter 3, the project is defended against a fundamental objection from Donald Davidson according to which conventions are in a sense not essential for there to be meaning.

In the subsequent chapters 4 to 9, theories of the three paradigms are critically discussed:

- Signaling Games: Theories of the first paradigm are a topic of active research in game theory today. According to them, language users are agents that either, as speakers, make observations upon which they send a signal or, as hearers, observe the signals upon which they react in typical ways. There are two standard interpretations of such theories. According to the rationalistic one, agents deliberate about which signal to send and how to react upon observing one. According to the alternative interpretation, agents are disposed to exhibit signaling behavior, however their dispositions are realized (typically they result from learning).
- Actual Language Relations: Theories of the second paradigm are considered to be the standard in analytic philosophy. As a building block they use linguistic theories of the kind linguists develop. Such a building block is related to the social practices of language use in a community.
- Evolutionary Theories: Theories of the third paradigm conceive of language as a bag of conventional behaviors which are evolutionary beneficial (in the wide sense of "evolution" which includes culture and biology). An important exponent of such a theory is Ruth Millikan.

In chapters 4 to 6, current theories of the first two paradigms are discussed (mostly under a rationalistic interpretation); they go back to David Lewis. Several problems are observed which are difficult to solve. Consequently, I plead for a theory of the third paradigm. In chapters 7 to 9, accounts of the third paradigm are discussed.

In chapter 7, Ruth Millikan's and Simon Huttegger's accounts are evaluated, the latter being an Evolutionary Signaling Games theory. Problems and limitations are observed. Among the problems is the lack of an adequate account of social norms.

In chapter 8, an account of social norms is developed as it is required for the "normativity of meaning" thesis. Thereby, one of the central questions of this thesis is answered, albeit not completely.

In chapter 9, an alternative conventionalist account is developed using Millikan's account of conventions, my account of social norms, and a new description of communication which allows us to assign meanings to words directly. The account features a unique combination of characteristics that make it comparably better than the alternative conventionalist accounts discussed in this thesis.

In chapter 10, the main claims of the dissertation are summarized together with a list of questions for future research.

Titles in the ILLC Dissertation Series:

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander van den Bosch**
Rationality in Discovery - a study of Logic, Cognition, Computation and Neuropharmacology

ILLC DS-2001-03: **Erik de Haas**
Logics For OO Information Systems: a Semantic Study of Object Orientation from a Categorial Substructural Perspective

ILLC DS-2001-04: **Rosalie Iemhoff**
Provability Logic and Admissible Rules

ILLC DS-2001-05: **Eva Hoogland**
Definability and Interpolation: Model-theoretic investigations

ILLC DS-2001-06: **Ronald de Wolf**
Quantum Computing and Communication Complexity

ILLC DS-2001-07: **Katsumi Sasaki**
Logics and Provability

ILLC DS-2001-08: **Allard Tamminga**
Belief Dynamics. (Epistemo)logical Investigations

ILLC DS-2001-09: **Gwen Kerdiles**
Saying It with Pictures: a Logical Landscape of Conceptual Graphs

ILLC DS-2001-10: **Marc Pauly**
Logic for Social Software

ILLC DS-2002-01: **Nikos Massios**
Decision-Theoretic Robotic Surveillance

ILLC DS-2002-02: **Marco Aiello**
Spatial Reasoning: Theory and Practice

ILLC DS-2002-03: **Yuri Engelhardt**
The Language of Graphics

ILLC DS-2002-04: **Willem Klaas van Dam**
On Quantum Computation Theory

ILLC DS-2002-05: **Rosella Gennari**
Mapping Inferences: Constraint Propagation and Diamond Satisfaction

ILLC DS-2002-06: **Ivar Vermeulen**
A Logical Approach to Competition in Industries

ILLC DS-2003-01: **Barteld Kooi**
Knowledge, chance, and change

ILLC DS-2003-02: **Elisabeth Catherine Brouwer**
Imagining Metaphors: Cognitive Representation in Interpretation and Understanding

ILLC DS-2003-03: **Juan Heguiabehere**
Building Logic Toolboxes

ILLC DS-2003-04: **Christof Monz**
From Document Retrieval to Question Answering

ILLC DS-2004-01: **Hein Philipp Röhrig**
Quantum Query Complexity and Distributed Computing

ILLC DS-2004-02: **Sebastian Brand**
Rule-based Constraint Propagation: Theory and Applications

ILLC DS-2004-03: **Boudewijn de Bruin**
Explaining Games. On the Logic of Game Theoretic Explanations

ILLC DS-2005-01: **Balder David ten Cate**
Model theory for extended modal languages

ILLC DS-2005-02: **Willem-Jan van Hoeve**
Operations Research Techniques in Constraint Programming

ILLC DS-2005-03: **Rosja Mastop**
What can you do? Imperative mood in Semantic Theory

ILLC DS-2005-04: **Anna Pilatova**
A User's Guide to Proper names: Their Pragmatics and Semanics

ILLC DS-2005-05: **Sieuwert van Otterloo**
A Strategic Analysis of Multi-agent Protocols

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder…": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics