1. Introduction

How many meanings does a modal expression such as ‘may’ have? One, says the Kratzerian orthodoxy in philosophy and much of linguistics (though one that yields different semantic values in different contexts). Many — or, more precisely: more than one — is the answer that we want to defend in this paper. Modals such as ‘may’, in other words, are ambiguous. It is often assumed that ambiguity should be minimised: where we can give a unified semantics for an expression, we should do so. We disagree, and we believe that the case of modal expressions can be seen as a model case for resisting the pull towards univocality. Crucial in the defence of our claim will be the observation that not all ambiguities are alike. There is homonymy, as in the case of ‘bank’; and there is polysemy, where the meanings of an ambiguous expression, while distinct, can be interestingly and systematically related to each other. Modal auxiliaries, we will claim, are polysemous. Since the standard account of these expressions explains what we take to be their polysemy as a case of context-sensitivity, we hope, by discussing our thesis, to shed some light on the under-explored relations between polysemy and context-sensitivity.

After some introductory remarks on modals in this section, we will examine the mechanisms of polysemy and context-sensitivity and provide criteria with which they can be held apart (Section 2). We will then apply the criteria to modal auxiliaries and show that the default hypothesis should be that they are polysemous, and not merely context-sensitive (Section 3). We will respond to arguments against modal ambiguity (and thus against polysemy) (Section 4). Finally, we will argue that modal polysemy has significant philosophical implications (Section 5). But first, let us take a closer look at modals and their standard analysis.

The main subjects of this paper are the modal auxiliaries ‘may’ and ‘can’. These are flexible expressions. On the one hand, speakers can use them to express different flavours of modality. For instance, ‘may’ in (1) could be used to express either epistemic or deontic modality:

(1) Holmes may travel to Paris.
If ‘may’ is used epistemically, (1) conveys that Holmes may possibly travel to Paris; and if it is used deontically, it conveys that Holmes is permitted to travel to Paris. In a similar manner, a speaker can use ‘can’ in (2) to express different flavours of modality:

(2) Holmes can play the violin.

Here, the possible modal flavours are either deontic or dynamic modality. On a deontic use of ‘can’, (2) conveys that Holmes is permitted to play the violin, while on a dynamic use, (2) conveys that Holmes is able to play the violin. ‘May’ and ‘can’ are thus flexible between modal flavours.

On the other hand, they exhibit flexibility even if we concentrate on a certain modal flavour. Consider several utterances of (1), in which ‘may’ is used in an epistemic way. These utterances can convey different things. One of them might convey that, for all Moriarty knows, Holmes may travel to Paris, while another might convey that, for all Watson knows, Holmes may travel to Paris. Different bodies of information are in play on different epistemic uses of ‘may’, and likewise different rules are relevant for different deontic uses. Next, imagine several utterances of (2) featuring a deontic ‘can’. These utterances might convey that Holmes is permitted to play the violin according to the hotel rules or according to the British laws. The same holds for dynamic ‘can’: (2) may be used to express, for instance, that Holmes has the ability to play the violin, or that there is an opportunity for him to do so. So ‘may’ and ‘can’ can be used to talk not only about different flavours of modality, but also about different possibilities within a certain flavour of modality. They are thus flexible in two different ways: between modal flavours and within modal flavours. This flexibility is typical for modals and can also be found with other modal auxiliaries such as ‘must’, ‘could’ and ‘should’ and semimodals such as ‘ought to’, ‘have to’ and ‘need to’. Our focus will be on ‘may’, ‘can’ and occasionally ‘must’, but the following considerations should apply to most or even all modal auxiliaries and semimodals that can be used to express several modal flavours. For brevity, we will henceforth simply use ‘modals’ to refer to modal auxiliaries.

How should the flexibility of modals be captured in their semantics? The most popular answer to this question takes it to be entirely a matter of semantic context-sensitivity. This contextualist semantics of modals was developed by Angelika Kratzer (1977, 1981, 1991) and is nowadays widely accepted in linguistics and the philosophy of language. To get a rough idea of how the contextualist approach works, let us briefly focus on ‘may’. Contextualists hold that the flexibility of ‘may’ between modal flavours is due to the fact that it functions as an existential quantifier that ranges over different kinds of possibilities in different contexts of utterance. For deontic uses of ‘may’, the domain of the quantifier contains deontic possibilities, while for epistemic uses, it contains epistemic possibilities. The contextualist would thus say that the semantic value of a certain epistemic use of (1) can be made perspicuous roughly as follows: there are some epistemic possibilities that are compatible with Holmes travelling to Paris. Which kinds of possibilities a modal quantifies over is determined by the context in which it is uttered. So contextualists say that ‘may’ is flexible between modal flavours because it is context-sensitive.

In order to account for the flexibility of ‘may’ within modal flavours, contextualists again invoke context-sensitivity: the domain over which ‘may’ quantifies does not only contain different kinds of possibilities in different contexts; it also contains different sets of possibilities (of a certain kind) in different contexts. For example, epistemic uses of ‘may’ quantify over different epistemic possibilities in different contexts. That is why different epistemic uses of (1) can convey different things. If (1) is used to convey that, for all Moriarty knows, Holmes may travel to Paris, then this is because its semantic value is (roughly) that there are some of Moriarty’s epistemic possibilities that are compatible with Holmes travelling to Paris. Again, the context determines how the domain of the quantifier is restricted, just as it determines which flavour this domain has.

While the contextualist view entails that the semantic values in context of modals vary in these two ways, it does not entail that modals bring two contextual arguments or parameters into sentences.
How many meanings for ‘may’? The case for modal polysemy

The aim of this paper is to argue that the first way in which modals are flexible, the flexibility they exhibit between modal flavours, is not a matter of context-sensitivity but is instead due to a certain kind of ambiguity, namely polysemy. Each modal does not have a single context-sensitive meaning, but rather several related meanings, each of which has a certain modal flavour. For example, ‘may’ has an epistemic meaning and a deontic meaning, and ‘can’ has a deontic meaning and a dynamic meaning. We do not wish to dispute that the flexibility of modals within a modal flavour is a case of context-sensitivity. While some theorists have argued that this flexibility should be accounted for by a pragmatic mechanism (Bach 2011, Braun 2012, 2013), we will follow, up until Section 5, the majority of theorists and accept that modals are context-sensitive in this respect (though we will not argue for this point and place no weight on it). So we do not wish to deny that modals are context-sensitive in some way or other; our point is that they are not merely context-sensitive. Modals are also polysemous.

What is it for an expression to be polysemous? And how can we distinguish polysemy from other kinds of ambiguity on the one hand, and from context-sensitivity on the other? These are the questions we will turn to in the following section.

2. Ambiguity and context-sensitivity

2.1 Ambiguity, homonymy and polysemy

Ambiguity, homonymy and polysemy are usually characterised as follows. An expression is ambiguous if it has multiple meanings. An expression is polysemous if it is ambiguous and its meanings are related to each other, and it is homonymous if it is ambiguous but its meanings are unrelated. Examples of homonymous expressions (and some of their unrelated meanings) include:

1. Throughout this paper, we will use logical form to denote the underlying semantic structure of an expression or a sentence. The logical form of an expression or a sentence may differ from its surface structure. For example, ‘ready’ can occur with just one argument at the level of surface structure (as in ‘Paul is ready’); but it presumably has two arguments at the level of logical form, as an agent can only be ready for something, and not merely ready.

2. This is a simplified outline of the contextualist story. Usually, such a story in which they occur. Rather, the standard view is that by fixing a certain set of possibilities (e.g. a set of epistemic possibilities), the context determines both the flavour of a modal and which possibilities (of that flavour) are at issue. This is evident in the logical form of contextualists ascribe to simple modal sentences like (1). This logical form consists of the following three parts: \([\text{Quantifier: Restrictor}] \ [\text{Scope}]\). The first position in this structure is occupied by an existential or universal quantifier (depending on whether the modal expresses a possibility or a necessity). The second position is filled by a function from worlds to sets of worlds; the output of this function is the domain of possibilities over which the modal quantifies, known as the modal base. Alternatively, one might say that the second position is filled simply by a proposition; the modal base in that case is the set of worlds at which the proposition holds. The third position is occupied by the proposition said to be compatible with all or some of the possibilities in the modal base; this proposition is sometimes called the prejacent. The logical form of (1) thus looks as follows:

\[
(1) \text{LF} \quad \text{[In some of the possibilities : [Holmes travels to Paris]. such that x holds]}
\]

\[
[\text{Quantifier: Restrictor}] \ [\text{Scope}].
\]

Only the second element in this structure is contextually variable. In any context, (1) features an existential quantifier and the prejacent that Holmes travels to Paris. But the modal base varies across contexts. It may be restricted by a certain epistemic proposition in one context, by a different epistemic proposition in a second context, and by a certain deontic proposition in a third context.

3. For example, by Elbourne (2011: Ch. 4) and by Hawthorne & Lepore (2011: 470–471).

4. By ‘expression’ we mean a (meaningful) string of letters. On this use, it is true
EMANUEL VIEBAHN & BARBARA VETTER

- ‘bank’ (financial institution/riverbank),
- ‘bark’ (part of tree/dog sound),
- ‘date’ (temporal indicator/a type of fruit),
- ‘lap’ (of a person/of a course) and
- ‘port’ (harbour/reduced wine).

Examples of polysemous expressions (and some of their related meanings) include:
- ‘book’ (abstract work/concrete copy),
- ‘date’ (temporal indicator/rendezvous),
- ‘healthy’ (as applying to animate objects/as applying to food or activities),
- ‘long’ (spatial distance/temporal distance),
- ‘mouth’ (part of an animal/of a river/of a bottle),
- ‘paper’ (material/newspaper/corporation issuing newspaper),
- ‘run’ (as a verb/as a noun),
- ‘since’ (causal relation/temporal relation) and
- ‘wood’ (small forest/material).

5. With these examples and the examples to follow, our aim is merely to provide rough approximations of some of the meanings; often there will be further meanings. The fact that ‘date’ appears in both lists of examples shows that some expressions are both homonymous and polysemous, with some unrelated meanings and some related ones. Another such expression is ‘bank’, which can be used as a verb (as in ‘to bank on someone’) and then has a meaning that is related to its financial-institution meaning.

PHILOSOPHERS’ IMPRINT

Vol. 16, No. 10 (June 2016)

How many meanings for ‘may’? The case for modal polysemy

Sometimes the difference between homonymy and polysemy is introduced with the metaphor of lexicon entries. A homonymous expression receives two separate lexicon entries, while a polysemous expression receives just one entry with several sub-entries.

This should provide an initial grip on homonymy and polysemy, but it leaves open a number of questions. What exactly is it for the different meanings of an expression to be ‘related’? And given that relatedness is a matter of degree (the meanings of ‘run’ appear to be more closely related than those of ‘mouth’, for instance), how closely do meanings have to be related in order to make an expression polysemous? Finally, how can we tell whether an ambiguous expression is homonymous or polysemous? We do not have worked-out answers to these questions (and since our main concern is with the difference between polysemy and context-sensitivity, we will not need them), but it is worth noting that there are a number of established patterns for spelling out the relatedness of meaning for polysemous expressions. These patterns of explanation can be used to distinguish homonymous expressions from polysemous ones. We will go through a few of them, not merely for the sake of illustration, but because these patterns will be useful in distinguishing between polysemy and context-sensitivity as well.

If we look at examples of polysemous expressions, certain typical relations between their respective meanings become apparent. Among these relations are the following:

Constitutive relations, which hold between objects and the stuff of which they are constituted: ‘wood’ as a collection of trees and the stuff out of which a tree is mostly constituted, ‘paper’ as the printing material or the newspaper that is printed on it, ‘duck’ as an animal or its meat, ‘cabbage’ as an individual vegetable or as cabbage-substance.
Causal relations, as they hold between producer and product: ‘(news)paper’ as the corporation or the physical copy produced by it, ‘milk’ as a liquid (noun) and the activity of expressing it (verb) or ‘cut’ as the activity or the result of cutting.


Two further kinds of relations are less easy to pin down, but no less common, and will be important in what follows:

Metaphorical extension: As John Lyons (1977: 552) notes, ‘the basic, or literal, meaning of ‘mouth’ is something like ‘aperture in the face (through which men and animals take food, breathe, emit vocal signs, etc.)’ and […] this meaning has given rise, by some discernible process of metaphorical or figurative extension, to the use of the same word in referring to other kinds of openings or apertures.’ Other examples of metaphorical extension include ‘long’, applied to spatial distance and, by metaphorical extension, to ‘distance’ in time; and ‘see’, used for visual perception and, by metaphorical extension to the realm of the mental, for intellectual understanding (‘I see what you mean’). (See Sweetser [1990] for more examples and a general account of such metaphorical extensions.)

Pragmatic strengthening: Elizabeth Traugott (1989) has argued that an expression with a given initial meaning may, by a gradual process, acquire a further meaning that is typically pragmatically implicated by its initial meaning. A standard example is ‘since’, which started out with a meaning for temporal succession (‘I have not heard from him since we last met’), often leading to a conversational implicature of causality (‘He has been unhappy since she left’). That implicature gradually became conventionalised and finally was established as a meaning of its own (‘I could not see him since it was so dark’). Similar developments are attested for ‘while’ (which started with a temporal meaning and then gained a meaning similar to ‘although’) and ‘observe’ (first meaning ‘perceive’, later also ‘remark’).

If the meanings of an expression exhibit the last two kinds of relations (and often also if they exhibit the relations we mentioned earlier), there is a ‘core’ or ‘initial’ meaning, from which further meanings are derived (temporal succession for ‘since’, spatial extension for ‘long’, the concrete token for ‘book’ and ‘newspaper’, and so forth). This core meaning has explanatory priority: it is because the expression has this meaning, and because the other meanings are in certain ways related to it, that the expression has its other meanings. The core meaning also has historical priority: it is the meaning with which an expression has started out, and from which other meanings have developed over time.

A straightforward way to establish whether a given ambiguous expression is homonymous or polysemous is to look for the kinds of relations we have discussed. Often it is easy to see whether the meanings of an expression stand in such relations, especially if the relations are constitutive, causal or instantiating. Relations of metaphorical extension or pragmatic strengthening can be harder to detect, but here we can turn to historical linguistics for help. If the etymology of an expression is known, it is possible to see whether its meanings have a common origin and arose through metaphorical extension or

6. See also Bybee et al. (1994: 196f).

7. This phenomenon was already noted by Aristotle, who argued that certain expressions exhibit a pros hen (‘towards one’) structure with a core meaning of explanatory priority. G.E.L. Owen, in an influential discussion of Aristotle’s metaphysics, has labelled this structure “focal meaning” (Owen 1960).
pragmatic strengthening. However, it is important to keep in mind that a common historical origin of several meanings cannot, by itself, provide a sufficient condition for polysemy. As Lyons (1977: 550) points out, a shared etymology is at best an imperfect guide to the polysemy/homonymy distinction, since even the meanings of homonymous expressions may be etymologically related. The two meanings of ‘port’ (harbour/reduced wine), for instance, are indirectly related in that both are derived from the Latin word ‘portus’, though by different routes (the wine is named after a city in Portugal, which in turn has its name derived from ‘portus’). Likewise, the meanings of ‘bank’ (financial institution/riverbank) have a common etymological origin in a Proto-Germanic expression for a shelf or a bench.\(^8\) Still, if the shared history of two meanings of an expression is fairly direct, and if one can in addition make out relations such as those of metaphorical extension or pragmatic strengthening, then it is likely that the expression is polysemous rather than homonymous. And if it is known that two meanings of an expression have distinct origins, then this is a strong indication for homonymy.\(^9\)

2.2 Polysemy and context-sensitivity

In our discussion of homonymy and polysemy, we talked as if expressions had just one kind of meaning (though possibly several exemplars of this kind of meaning). The kind of meaning we were concerned with was standing meaning: the kind of meaning that would appear in a dictionary. Such talk is fine as long as context-sensitivity stays out of the picture. But once we deal with semantic context-sensitivity, we have to distinguish between the standing meaning of an expression and the semantic value it has on a certain use or in a certain context. For the mechanism of semantic context-sensitivity is based on precisely this distinction: it is said to involve an invariable standing meaning that leads to variable semantic values. In what follows, we will use meaning to refer to the stable, standing meaning or meanings of an expression, and semantic value to refer to the value it receives in a given context.\(^10\)

If an expression is neither context-sensitive nor ambiguous, it has one standing meaning, which is identical to its semantic value (or is a constant function from contexts to a certain semantic value). If an expression is ambiguous but not context-sensitive, it has several standing meanings, each of which is identical with a single semantic value (or, again, is a constant function from contexts to a certain semantic value). How might one then characterise context-sensitive expressions? As a first attempt, one might say that an expression is semantically context-sensitive if it has a single standing meaning that determines different semantic values in different contexts of utterance.\(^11\) This characterisation applies to expressions that are merely context-sensitive, such as ‘I’ and ‘today’. But not all context-sensitive expressions are merely context-sensitive: an expression can be both context-sensitive and ambiguous. For example, it seems plausible to say that ‘long’ and ‘healthy’ are context-sensitive as well as polysemous. Consider ‘long’: this expression has a spatial and a temporal meaning, but each of these meanings appears to be context-sensitive. It can be true in one context but false in another to call one and the same entity (spatially) ‘long’, and the same goes for its temporal meaning. So there are some apparently context-sensitive expressions that have more than one standing meaning and thus do not fulfil the criterion introduced above. We can avoid this problem by ascribing context-sensitivity not to expressions but rather to meanings: a meaning is context-sensitive if it determines different semantic values in different contexts of utterance. An expression can then

---

9. A clear case of unrelated meanings is that of ‘date’: the OED specifies that its fruit meaning probably derives from the ancient Greek ‘δάκτυλος’ (‘finger’), possibly because of the elongated shape of a date fruit, while its temporal meaning stems from the Latin ‘datum’ (‘given’).
10. In Kaplanian terms, the distinction is one between the character and the content of an expression. Cf. Kaplan (1989a, 1989b).
11. Henceforth we will use ‘context-sensitivity’ instead of the more cumbersome ‘semantic context-sensitivity’.
be said to be context-sensitive if it has at least one context-sensitive meaning, and it can be said to be merely context-sensitive if it has exactly one context-sensitive meaning. This approach allows us to characterise ‘I’ and ‘today’ as merely context-sensitive expressions, and ‘healthy’ and ‘long’ as expressions that are both polysemous and context-sensitive. For the purposes of this paper this characterisation should suffice. It may help to display the three cases of mere polysemy, polysemy cum context-sensitivity and mere context-sensitivity schematically as in Fig. 1:

Fig. 1: Polysemy and context-sensitivity

Now let us consider methods with which we can decide whether an expression is context-sensitive, polysemous or both (ignoring the uninteresting case where it is clearly neither polysemous nor context-sensitive). Distinguishing between these three kinds of expressions is complicated by the fact that expressions of each kind have commonalities with expressions of the other two kinds. In what follows, we will continue to call an expression merely context-sensitive if it is context-sensitive but not polysemous, and merely polysemous if it is polysemous but not context-sensitive. Even expressions that are merely context-sensitive and expressions that are merely polysemous have some things in common: both kinds of expressions have different semantic values in different contexts of utterance; in other words: they have several candidate semantic values. The semantic values of a merely polysemous expression vary across contexts because different meanings of that expression are operative in different contexts, while the semantic values of a merely context-sensitive expression vary across contexts because its (single) meaning determines different semantic values in different contexts. So we have two different mechanisms that lead to results that are similar in kind (namely different semantic values in different contexts, and thus several candidate semantic values), and it can be hard to tell which mechanism is at work in a given case. Still, we think that there are a number of methods and criteria, which can jointly help us to decide whether an expression belongs to the group of polysemous expressions, to the group of context-sensitive expressions or to both.

1. Linguistic intuitions. As a starting point, we can look at an expression and ask whether, intuitively, it has several meanings or not. Each of the following expressions can have different semantic values in different contexts of utterance: ‘book’, ‘wood’, ‘healthy’, ‘long’, ‘I’, ‘today’ and ‘tall’.\(^\text{12}\) But while it seems plausible to say that ‘book’, ‘wood’,

12. There is an on-going debate about whether ‘tall’ is context-sensitive or not (and thus about whether it has different semantic values in different contexts of utterance). We cannot enter this debate, and will simply assume for this paper that ‘tall’ is context-sensitive. See Stanley (2002), Kennedy (2007) and others for arguments in favour of context-sensitivity, and Borg (2012) and
‘healthy’ and ‘long’ have several meanings, this does not seem plausible for ‘T’ and ‘today’. This gives us a reason to think that the former expressions are polysemous and the latter merely context-sensitive.

The use of linguistic intuition is, of course, limited. For one thing, we may not always have strong intuitions about whether an expression has one or several meanings. Take ‘tall’, for example. At first view this expression has just one meaning, having to do with the height of an object. But one can also use ‘tall’ in phrases such as ‘tall task’ or ‘tall order’. Are these further meanings of ‘tall’ or merely non-literal uses? Intuitions will probably not settle such cases. A second limitation is that we cannot use this method to distinguish merely polysemous expressions from expressions that are polysemous as well as context-sensitive. However, linguistic intuitions can serve as a valuable starting point for applying the following four defeasible but plausible criteria.

(2) Numbers of candidate semantic values. Context-sensitive expressions, it seems, have many more candidate semantic values than merely polysemous expressions. Compare the (presumably) context-sensitive ‘T’, ‘today’ and ‘tall’ with the (presumably) merely polysemous ‘book’ and ‘wood’. ‘T’ has as many candidate semantic values as there are speakers of English, and ‘today’ as many as there are days. Both expressions are even trumped by ‘tall’, which has as many candidate semantic values as there are standards of tallness. By contrast, ‘book’ has to get by with far fewer candidate semantic values: one for an abstract work of literature, a second for a concrete entity made of paper, a third for a subdivision in a treatise, and possibly a few more. Likewise, ‘wood’ has only a handful of candidate semantic values. Such a difference in the number of candidate semantic values seems to exist whenever we compare a clear example of a context-sensitive expression with a clear example of a merely polysemous expression. So we can adopt a criterion according to which an expression is likely to be context-sensitive if it has many candidate semantic values, and likely to be merely polysemous if it has few candidate semantic values.

The first direction of this criterion can possibly be made even stronger with the help of the following consideration: if an expression has infinitely (or near infinitely) many candidate semantic values, it *has to* be context-sensitive, for otherwise we would have to implausibly assume that there are expressions for which we store infinitely many meanings in our finite brains. So we have very good reasons to take an expression with infinitely or very many candidate semantic values to be context-sensitive, and not merely polysemous (or homonymous). The opposite direction of the criterion cannot be reinforced in the same way: in principle, there could be context-sensitive expressions with only a few candidate semantic values. But it still appears to be the case that only the clear examples of merely polysemous expressions have few candidate semantic values, so if an expression has a small number of candidate semantic values, that is indicative of its mere polysemey.

This criterion does not help to distinguish between merely context-sensitive expressions and expressions that are both context-sensitive and polysemous. But in conjunction with criterion (1), which does distinguish between these two groups of expressions, we can make a first stab at a three-way distinction. Taken together, the criteria suggest that ‘book’ and ‘wood’ are merely polysemous, and that ‘T’ are ‘today’ are merely context-sensitive. They place ‘healthy’ and ‘long’ (and possibly also ‘tall’) in the middle group of context-sensitive and polysemous expressions, as these expressions each seem to have very many candidate semantic values, corresponding to different degrees of health and length (and tallness), as well as a smaller number of distinct meanings, corresponding to the spatial and temporal reading of ‘long’, etc.

(3) Clusters of candidate semantic values. We have seen that context-sensitive expressions have a large number of candidate semantic values. In order to distinguish between merely context-sensitive expressions and those that are both context-sensitive and polysemous, we...
can observe how the candidate semantic values of an expression are distributed. In particular, we can look out for clusters of candidate semantic values. It is hard to formally define what a cluster of candidate semantic values is, but we can introduce it by way of example: ‘long’ seems to have two (non-overlapping) clusters of candidate semantic values, one consisting of spatial distances, the other of temporal distances (durations). Likewise, ‘healthy’ has (at least) two clusters of candidate semantic values, one consisting, roughly, of various degrees of flourishing in animate objects, and the second of properties possessed by food or activities. On the other hand, the candidate semantic values of ‘I’ and ‘today’ do not come in clusters; their candidate semantic values form a set that is relatively homogeneous and not naturally organised into disjoint subsets. As this pattern repeats itself for other expressions with many candidate semantic values, we can say that an expression is likely to be merely context-sensitive if it has only one cluster of candidate semantic values and that it is likely to be both context-sensitive and polysemous if it has several clusters of candidate semantic values.

(4) Relations among candidate semantic values. In the previous section, we have seen various ways in which the meanings of polysemous expressions may be related. We have also seen that these relations often privilege one meaning, the core meaning, which has explanatory and historical priority over other meanings. In order to detect polysemies, we can look for such typical relations and for explanatory and historical priorities among the candidate semantic values of an expression. The candidate semantic values of merely context-sensitive expressions do not appear to exhibit these kinds of structures. No candidate semantic value of ‘I’ or ‘today’ takes priority, qua semantic value of ‘I’, over the others, temporally or explanatorily. The different semantic values are not typically related by metaphor, pragmatic strengthening or any of the other relations that we have listed for polysemous meanings. Nor should this come as a surprise. With mere context-sensitivity, we have a single meaning that gives us a function from contexts to semantic values. Different semantic values are determined by different contexts of utterance; but all semantic values that could be determined in context, i.e. all candidate semantic values, can plausibly be said to be on a par.

Of course a context-sensitive expression can have some prior (or core) candidate semantic values if it is also polysemous. In that case, certain clusters of candidate semantic values have explanatory and historical priority over others, or are related to others by metaphor, pragmatic strengthening or other typical relations. For example, the spatial candidate semantic values of ‘long’ are explanatorily and historically prior to its temporal candidate semantic values, and the latter arose as a metaphorical extension of the former.

So the current criterion tells us that if, among an expression’s candidate semantic values, we find some of the typical structures of polysemy sketched above, this is good evidence in favour of polysemy and against mere context-sensitivity. Historical linguistics is useful in applying this criterion, for it can help us identify prior candidate semantic values and paths by which the other candidate semantic values have developed.

(5) Logical form. The idea of our final criterion is that if, among the candidate semantic values of an expression, some occur only in one kind of construction at the level of logical form, while others occur only in a different kind of construction at the level of logical form, then that is evidence that the expression is polysemous, and not merely context-sensitive. The reason is simple: merely context-sensitive expressions have only one meaning, but a meaning that must be appropriate to the logical type of the expression. Thus the meaning of ‘T’ is standardly taken to be a function from contexts of utterance to individuals, viz. the speakers of the respective context. And the meaning of ‘today’ is taken to be a function from contexts to days. An expression that can function either as a verb or as a noun, say, could not be assigned this kind of uniform meaning. As a verb it would have to be assigned a function from contexts to a predicate-type intension (properties or functions from possible worlds to sets of individuals); as a noun, a function from contexts to a noun-type intension (individuals
that an expression is context-sensitive in the relevant respect. This criterion is based on the view that, wherever possible, we should strive for unity in semantic theories: if we have an expression with several candidate semantic values, ascribing one context-sensitive meaning is preferable to ascribing several non-context-sensitive ones. Here is a recent formulation of such a maxim, which is explicitly aimed at the case of modals:

Other things equal, linguists rightly prefer more to less unified semantic theories. Such theories are simpler and so allow for more plausible explanations for how speakers are able to learn expressions and to competently use and understand their uses in new contexts. In the modal case, the ideal would be a theory that gave a single, unified semantic treatment of all of our modal expressions, not only the deontic modals or the deontic and the epistemic modals, but uncontroversially circumstance-relative ones (such as ability modals) as well. (Dowell 2013: 154)

Such a criterion appears to speak strongly against our present argument. For, it seems, we can give a unified, context-sensitive mechanism in the case of modal flavours; that is precisely what Kratzer and others have done. We will not argue that these mechanisms fail to deliver the right results. So what is wrong with the criterion?

The problem lies with the background assumption that unity, and thus context-sensitivity, should be maximised in semantic theories. While it might be easier for speakers to learn expressions governed by simpler, more unified theories, this does not mean that speakers in fact learn and use expressions that conform to such theories. In order to arrive at an accurate theory, semanticists have to take into account current and historical usage. A maxim of unity might then be brought in if two theories are empirically equivalent. But in the next section, we

will employ our criteria to show that standard contextualist theories of modals are not on a par with theories incorporating polysemy. We thus have a first reason not to maximise unity and context-sensitivity. Furthermore, it is not so clear that unity should be maximised even if theories are empirically equivalent. The alleged advantage of unity and context-sensitivity is to make it easier to learn and use expressions. But there are other ways in which expressions can be easily learnable. What seems to matter, for instance, is that (most) expressions with multiple candidate semantic values display a certain amount of systematicity. Systematicity, however, cannot be found only with context-sensitivity, but also with polysemy. After all, the distinct meanings (and candidate semantic values) of polysemous expressions are usually systematically related, as we have shown above. And even if polysemous expressions display slightly less systematicity than context-sensitive ones, we have seen that some such mechanisms must be posited anyway to account for the uncontroverted cases of polysemy. Rather than maximise context-sensitivity at the expense of polysemy, we should accept that there are two ways for an expression to have systematically correlated candidate semantic values, and for any given expression we should check whether the kind of correlation that it shows “fits in” with the paradigm examples of polysemy, or those of context-sensitivity. Our proposed five criteria do exactly that. It is time now to apply them.

3. The polysemy of modals

In this section, we will argue that polysemy is the default view in understanding the different modal flavours of expressions like ‘may’ and ‘can’. In some ways, this will be expected even by orthodox Kratzerians: after all, Kratzer’s view is taken to be progress because it is an improvement on the prima facie view. We disagree with that verdict, and will justify our disagreement in two prongs. In this section, we want to show that prima facie appearances speak in favour of the default view; in the next section, we will respond to arguments against the default view. The argument of this section will proceed, unsurprisingly, by applying the five criteria that we have developed in the previous section.

1) Linguistic intuitions. Intuitively, most modals do have several meanings, which correspond to different modal flavours: epistemic meanings which concern compatibility with what we know; deontic meanings which concern right and wrong, goals and preferences; and dynamic meanings which concern the abilities and dispositions of agents and objects. In particular, there is a strong intuition that ‘may’ has distinct epistemic and deontic meanings, and that ‘can’ has distinct deontic and dynamic meanings (as evidenced by everyday disputes about whether the deontic usage of ‘can’ is correct).

While the recent focus has been on the commonalities between different uses of modals, theorists have previously acknowledged the intuitions differences between their meanings. Lyons, for instance, states that it “has long been recognized that most of the sentences containing such [modals] as ‘must’ and ‘may’ are ambiguous” (1977: 791). And according to Palmer, “both MAY and MUST are used in two quite different ways, and this justifies the distinction between epistemic and deontic modality” (1990: 35). These statements are in line with our intuitive verdict that modals have several meanings, so the current criterion provides a first push towards modal polysemy. However, the intuitive verdict requires further support, and it indicates only whether an expression is polysemous, and not whether it is also context-sensitive. Once the four further criteria from Section 2.2 are applied, we will see that the intuitive verdict of polysemy is confirmed, and a view of modals as polysemous as well as context-sensitive becomes plausible.

2) Numbers and (3) clusters of semantic values. Most theorists hold that modals have a plethora, perhaps an infinity, of candidate semantic values. That seems right: many bodies of information can be in play on epistemic uses of ‘may’, and many sets of rules can matter for deontic uses of ‘can’. Our criterion (2) says that such large numbers of candidate semantic values are indicative of context-sensitivity. But the candidate semantic values of modals are grouped into clusters: ‘may’ has a cluster of epistemic candidate semantic values and a cluster of deontic
ones; ‘can’ has a deontic cluster and a dynamic cluster. According to criterion (3), this is a sign for polysemy. So criteria (2) and (3) together speak in favour of polysemy cum context-sensitivity.

4. Relations among candidate semantic values. We have argued above that if an expression exhibits one (or more) of the typical relational structures of polysemous expressions, and in particular if it has a ‘core’ meaning from which its other meanings can be derived, historically and explanatorily, then we have good reason to believe that the expression is polysemous rather than context-sensitive. Historical linguistics can provide clues here; and accordingly we will start with a look at the history of modal words.

Cross-linguistic surveys, in particular Bybee et al. (1994; see also van der Auwera & Plungian 1998), have shown that there is a robust pattern to the development of polyfunctional modals. A modal is polyfunctional if it can express several flavours of modality, so ‘may’ and ‘can’ are polyfunctional. Bybee et al. have provided strong evidence that non-epistemic meanings of modals (which are often grouped together under the label root meanings) develop first and epistemic meanings only much later across a large range of different languages (Bybee et al. used a representative sample of seventy-six languages in their survey; see pp. 27–32). Typically, the development starts with a main (non-modal) verb turning into a modal either of ability or of obligation. The starting points for the ability modals are often verbs of knowledge (as in ‘can’, which is related to ‘know’) or physical power (as in ‘may’, which is related to the noun ‘might’); the starting point for the obligation modals are often verbs that express obligation or the copula (as in the English ‘I am to go’). These root meanings develop into other root meanings — from the ability sense, for instance, Bybee et al. trace a development to “root possibility” (meaning, roughly, opportunity) and from there to permission — and finally into the epistemic meaning. Thus ‘can’ and ‘may’ both started out as expressions of ability, then developed a sense of permission (192–194). From there their paths diverge: ‘may’ has further developed to express epistemic possibility, and has lost the ability sense, while ‘can’ has retained the ability sense and acquired an epistemic sense only in very few contexts. ‘Must’, on the other hand, has developed from a modal expressing obligation into one that also expresses “inferred certainty” (195); and similarly it goes for other polyfunctional modals.

What does this show? First, modals do not come to express deontic, dynamic and epistemic modality “all at once”. This we have seen to be typical of polysemy rather than context-sensitivity: the different meanings of ‘long’, ‘mouth’ or ‘paper’, too, have developed over time, and in this respect contrast with merely context-sensitive expressions such as ‘I’ and ‘today’. Second, like many polysemous expressions, and unlike typical context-sensitive expressions, modals have (or used to have) a “core” meaning, which is in each case a particular root meaning: ability for ‘may’ and ‘can’, obligation for ‘must’. These core meanings provide the starting point for the historical development. This historical priority strongly suggests that there is explanatory priority as well: modals have certain epistemic meanings because they had certain root meanings in the first place. And indeed, the relation between these ‘core’ meanings and the derived (in particular, the epistemic) meanings can be classified in one (or both) of (at least) two ways that are typical of polysemy: as metaphorical extension, or as pragmatic strengthening.

The metaphorical approach has been defended by Eve Sweetser (1990). Sweetser’s main claim is that a great deal of (synchronic) polysemy and diachronic semantic change can be accounted for in terms of a cognitive operation that she calls metaphor, and which connects different domains. Take the polysemous ‘long’: there is a general metaphorical connection from the domain of space to that of time, by which time intervals are thought of as being like distances in space, and our existence in time as akin to a travel through space. The polysemy of ‘long’ is just one aspect of this metaphorical connection. Another pervasive area for metaphor is the use of perception words (such as ‘see’) to denote, in addition to physical perception, mental acts or states (‘I see’, in the sense of ‘I understand’; see Sweetser 1990: Ch. 2). For such metaphorical mappings it is not required that the two phenomena — the original and the metaphorical meaning — are in any
interesting sense species of some common genus; in fact, metaphor is characterized instead by the fact that the different meanings belong to different domains — space and time, or external perception and mental operations. What is required, Sweetser argues, is that there be certain structural commonalities between the different meanings. Again, it is easy to see this for the case of ‘long’: long things in space and long stretches of time share a number of structural features, such as the divisibility into many shorter items of the same kind (spatially extended objects, or stretches of time).  

How does this work for modals? Sweetser argues that the transition from root to epistemic modality is part of the general trend, also seen in the polysemy of ‘see’ mentioned above, to use the language of the external world to apply to the internal mental world, which is metaphorically structured as parallel to that external world. Thus we view our reasoning processes as being subject to compulsions, obligations, and other modalities, just as our real-world actions are subject to modalities of the same sort. (1990: 50)

The common structure, according to Sweetser, of root and epistemic modals is that of “forces” and “barriers”: ‘must’ indicates “a compelling force directing the subject towards an act”, while ‘may’ indicates “a potential but absent barrier” (52). Forces and barriers may be real (in the dynamic reading), normative or social (in the deontic reading), or epistemic: “The meaning of epistemic may would thus be that there is no barrier to the speaker’s process of reasoning from the available premises to the conclusion expressed in the sentence qualified by may” (59). Thus, while the deontic reading of our sentence

expresses the absence of social barriers for Holmes in travelling to Paris, its epistemic reading expresses the absence of any rational barriers for the speaker in believing that Holmes will travel to Paris. Forces and barriers, however, cannot be used to formulate uniform semantics for modals, since the different modal flavours are not literally, but only metaphorically, a matter of forces and barriers.

An alternative explanation in terms of pragmatic strengthening is offered by Traugott (1989). As we have briefly sketched above, this type of explanation posits a mechanism by which an expression comes to carry an implicature, which subsequently develops into a meaning of its own. (Recall the case of ‘since’, first expressing temporal succession and later a causal relation.) Similarly, the idea is, expressions of root modality first conversationally implicate epistemic possibility. When we ascribe abilities, it is typically because we expect them to be exercised at some point; and there is little point in expressing permission when we already know that the permitted action will not be performed. (Like any implicature, these can be cancelled: witness ‘Holmes can travel to Paris, but I know he won’t.’) On the present model, this conversational implicature first becomes a conventionalised implicature accompanying the (root) meaning of modals such as ‘can’ or ‘may’, and then develops into an independent, separate meaning, making the modals polysemous. Bybee et al. adduce a number of examples for the intermediate stage, where modals express root and epistemic meaning simultaneously, from Old English (see Bybee et al. 1994: 197ff.).

The two explanations are by no means mutually exclusive: each might apply to some cases but not others. Thus Bybee et al. (1994) argue for an implicature-based model in the case of ‘can’ and ‘may’, but appeal to metaphor for ‘must’ (cf. p. 200).

The historical results, together with the candidate explanatory hypotheses, constitute further evidence in favour of the polysemy thesis: modals exhibit just the kind of semantic structure that we should

---

16. This leaves a great many questions open: What individuates domains? And what makes shared features structural features? We do not have the space to go into Sweetser’s theory in any more detail, but we hope that the brief remarks made here are enough to give an idea of the relation of metaphorical extension that might hold between several meanings of polysemous expressions.
expect of a polysemous expression, but not of a merely context-sensitive one.\textsuperscript{17}

(5) Logical form. As Paul Portner (2009: 143) says, a “significant theme in the syntax literature is that different semantic categories of modals are located in different positions in syntactic structure. The most basic claim of this kind is that epistemic modals reside higher in the tree than non-epistemic ones.” We can justify and explain this common claim by looking at the respective scope of epistemic and non-epistemic modals.

To begin with, here is a pair of sentences that illustrates, on the most natural readings, the different scoping behaviour of root and epistemic modals with respect to tense:

(3) Mary had to take the train to go to Paris. (Deontic: In the past, it was necessary for Mary to take the train.)

(4) Mary had to be home (at the time of the crime). (Epistemic: It is necessary, given what we know now, that in the past Mary was home.)\textsuperscript{18}

By the transitivity of the “scope above” relation, it follows that epistemic modals must scope above root modals. For illustration, note that sentences (5) and (6) have only readings where the epistemic modals take scope over the root ones, independently of the word order:

(5) Joan may have to help you with that. (epistemic may > deontic have to)

(6) Mary can probably come tonight. (epistemic probably > dynamic/deontic can)

Moreover, where the surface syntax determines the relative scope of the modals, as in (5) (but not (6)), the only available reading is one where the outer modal is epistemic and the inner deontic, despite the polyfunctionality of both modals: there is no reading of (5) on which it says that Joan is allowed to be such that it is epistemically necessary for her to help the addressee.\textsuperscript{19}

How are we to explain these different scoping behaviours between epistemic modals on the one hand and root (that is, dynamic and deontic) modals on the other? A widely (albeit not universally) accepted conclusion from these observations in linguistics is that epistemic modals generally scope over whole sentences or clauses (including tense, aspect and any root modals), while root modals tend to scope just above the verb phrase (and hence under tense, aspect and any epistemic modals).\textsuperscript{20} Schematically, this can be seen in Fig. 2: the tree on the left-hand side gives the position of an epistemic modal, that on the right-hand side the position of a root modal:

\textsuperscript{17} Indeed, so far as we can tell, historical linguists seem to take it for granted that modals are polysemous. Thus Traugott (1989: 33) formulates a constraint on theories of semantic change that “[i]t seems that the semantic theory underlying a theory of semantic change must allow for polysemy”, and Bybee et al. (1994: 204) refer to “the polysemy of English may” as an established fact. However, since neither Traugott nor Bybee et al. seem to consider the alternative between polysemy and context-sensitivity (neither of them even mentions Kratzer), these remarks are not of much use for our present purposes. Sweetser (1990: 56) briefly addresses Kratzer’s work and rejects it as “subsum[ing] the root meanings under very general epistemic readings”, and thereby “fail[ing] to motivate the attested historical and developmental progression from root to epistemic, rather than in the other direction”. Sweetser’s departure from Kratzerian semantics is, however, much more radical than ours: she replaces truth-conditional semantics with what she calls a “cognitively based theory which takes … human perception and understanding of the world to be the basis for the structure of human language” (Sweetser 1990: 2). In appealing to her explanation by metaphorical extension, we do not intend to follow her in her larger project. On the contrary, we use her explanation, which is intended as a semantic one in her (cognitive) sense, precisely to show that we need not assume a single meaning (in our sense) to explain the relations between the candidate semantic values of modals. Sweetser’s metaphorical relations are thus emphatically not part of the semantics of modals on our view.

\textsuperscript{18} These examples are taken from Hacquard (2010: 87).

\textsuperscript{19} Nauze (2008: 175–177) uses similar data to argue for a view on which modals are polysemous.

\textsuperscript{20} See Brennan (1993), Hacquard (2011) and the references therein.
How many meanings for ‘may’? The case for modal polysemy

We have already seen that the difference does go hand in hand with an intuitive difference in meaning (our first criterion). Hence our fifth criterion, too, provides evidence in favour of the polysemy thesis.

We would like to note one limitation of this fifth criterion, which is shared to some extent by the fourth (concerning relations among the candidate semantic values). The syntactic differences that we have sketched hold between epistemic modality on the one hand and root modality — including at least deontic and dynamic modality — on the other. (To complicate matters, some linguists have argued that deontic modality is divided into a predicate-modifying ‘ought-to-do’ and a sentence-modifying ‘ought-to-be’ use; see Brennan 1993, Hacquard 2006, Portner 2009 and also Schroeder 2011.) It therefore favours a polysemy approach only between epistemic and root modality, and has no bearing on the distinction between deontic and dynamic modality, which intuitively (our first criterion) is a distinction between different meanings as well.

Nor is this observation restricted to English or related languages: Guglielmo Cinque (1999) has argued that this hierarchy holds across a large set of languages, including English, Bosnian/Serbo-Croatian, Hebrew and Chinese. The foregoing suggests that our sentence

(1) Holmes may travel to Paris

must be construed differently on two different readings. The idea is that (1) in its epistemic reading is about the proposition that Holmes travels to Paris; in particular, a given utterance of (1) tells us something about the speaker’s attitude towards that proposition. On its deontic reading, (1) is more directly about its subject, Holmes: it is used to attribute to Holmes himself the property of being permitted to travel to Paris. Similarly, ‘Holmes can travel to Paris’ attributes to Holmes himself the property of being able or the property of being allowed to travel to Paris.

We argued above, in Section 2.3, that such differences in the positions that an expression has in logical form are indicative of polysemy, rather than (mere) context-sensitivity. This seemed especially clear when the positions were as different as those of noun and verb (‘run’ as noun/as verb), or when the different positions in logical form went hand in hand with intuitively clear differences in meaning (‘wood’ as small forest/material). The difference between a predicate modifier and a sentence modifier may not seem as deep as that between a noun and a predicate, though we submit that it may well be deeper than that between count nouns and mass nouns. But we have already seen that the difference does go hand in hand with an intuitive difference in meaning (our first criterion). Hence our fifth criterion, too, provides evidence in favour of the polysemy thesis.

We do not claim that the Kratzerian framework cannot accommodate the data we have adduced. The framework is an extremely flexible one, and Virginia Brennan (1993) has indeed provided a variation on Kratzer’s theory that is aimed at incorporating the differences of logical form. It is not clear to us whether Brennan’s theory does not make modals polysemous after all, though she does not seem to think it does. In any case, our claim here, as above, is not that we cannot construe modals as merely context-sensitive, but that it is better to think of them as polysemous. Hacquard (2010, 2011) explicitly accepts that epistemic and root uses of modals have different logical forms, yet maintains that the modals are still univocal and merely context-sensitive. Again, we are not convinced that her account is a univocal one; and in any case, we have seen above that modals would then constitute a rather isolated and unusual case of context-sensitivity. In the absence of convincing arguments for univocality (see the next section), there is no reason not to accept what the data about logical form suggest: that modals are polysemous, rather than merely context-sensitive.

22. Cinque’s own hierarchy is much more complex than that depicted in Fig. 2, and his book is primarily concerned with adverbs, but he considers other expressions of tense, aspect and modality as well.
We believe that the evidence in total still suggests (at least) a threefold polysemy between epistemic, deontic and dynamic modal meanings. We would be almost content, however, if we had convinced the reader only that there is some, at least twofold, polysemy in modals.24

4. Arguments against the ambiguity of modals

In this section, we will discuss three arguments against the view that modals are polysemous. The first two arguments stem from Kratzer’s influential ‘What must and can must and can mean’ (1977) and may be labelled the explosion-of-meanings argument and the common-kernel-of-meaning argument. These arguments are aimed at ambiguity accounts of modals in general, and not specifically at polysemy accounts. But insofar as they apply to ambiguity in general, they will apply to polysemy in particular; and in any case they still surface in the current debate and have been taken to apply to ambiguity accounts of different kinds, including those positing polysemy (e.g. by Schaffer 2011). The third argument, which we will call the cross-linguistic argument, has been employed frequently in recent years to argue against ambiguity accounts in general and polysemy accounts specifically (e.g. by Hacquard 2011). It is the strongest argument against polysemy, so we will discuss it in detail. However, we will conclude that this argument provides nor more reason to abandon a polysemy account than the other two arguments.

24. See Nauze (2008: Ch. 4) for further arguments for modal polysemy. Unfortunately, we do not have the space here to discuss Nauze’s arguments for polysemy or his account of modals (which is a kind of update semantics). But we want to point out that Nauze approaches the debate from a linguistic, typological perspective, and that his case for polysemy overlaps with ours only with respect to the argument from logical form. If successful, Nauze’s arguments and our more theoretical, philosophical arguments for polysemy can nicely complement each other.

4.1 The explosion-of-meanings argument

Kratzer’s first argument against modal ambiguity is based on the worry that opening the door to ambiguity leads to an explosion of meanings.25 This worry is voiced in the following passage:

How many kinds of ‘must’ do we have to distinguish? How many deontic tones? How many epistemic tones? How many dispositional tones? And how many preferential ones? Obviously many, many of each group. We do not just refer to duties. We refer to duties of different kinds. To different duties different persons have towards different persons at different times. We do not simply refer to a bit of knowledge or information — once and for ever the same. We refer to different kinds of knowledge or information in different situations. […] All this leaves us with many different ‘must’s and ‘can’s. (1977: 339)

Kratzer appears to have an account of modals in mind that analyses both ways in which modals are flexible (between modal flavours and within modal flavours) in terms of ambiguity. Such an account would indeed require modals to have a huge number of meanings, and could rightly be criticised for this commitment. But other ambiguity accounts of modals do not lead to such an explosion of meanings. In particular, our account, on which only the flexibility between modal flavours is analysed as a kind of ambiguity, does not posit many meanings for modals. Rather, it posits just a handful of ‘can’s and ‘may’s, one for each flavour of modality. Kratzer’s worry is unfounded with respect to such an account and in general with respect to accounts that capture only the flexibility between modal flavours with the help of ambiguity, so we can set it aside for now.

4.2 The common-kernel-of-meaning argument

Kratzer’s second argument against an ambiguity account of modals does apply to accounts positing only a limited amount of ambiguity. It is based on her “feeling that there is something in the meaning [of different occurrences of modals] which stays invariant” (1977: 340). The common kernel of meaning Kratzer sees is that modals are quantifiers that take two arguments, one for a proposition restricting the modal base and one for a proposition that is its prejacent. Given that any occurrence of a modal shares this kernel of meaning, her argument continues, it would be strange to posit multiple meanings for some or all modals.26

We agree that different occurrences of modals appear to have something in common. But we do not agree that this speaks against every kind of ambiguity account of modals.

First, a common kernel of meaning is not a sufficient criterion for sameness of meaning. Expressions with distinct meanings can have a common kernel of meaning. For example, any reasonable semantic theory will provide a common kernel of meaning for the pronouns ‘he’ and ‘she’, but no such theory should entail that these two pronouns have the same meaning. The same holds for modals: Kratzer would say that occurrences of ‘might’ have a meaning that features an existential quantifier with two arguments, and that the same is the case for occurrences of ‘can’. But it would be strange to deduce from this observation that these two expressions have the same meaning, even for Kratzerian semantics; they certainly do not appear to be synonymous. And if occurrences of distinct expressions can have a common kernel of meaning without having the same meaning, this should equally be the case for different occurrences of a single expression.

Second, the common kernel of meaning Kratzer observes in modals fits in naturally with a polysemy account. Kratzer sees commonalities between different occurrences of a modal because of structural similarities between their meanings: each occurrence has a meaning that features a quantifier that takes two arguments. Such a common structure might be surprising if modals were homonymous. But it is far from surprising on a polysemy account of modals, which entails that the different meanings of a modal are related. Friends of polysemy can take the common structure of meanings of, say, epistemic and deontic uses of ‘may’ to be one of the aspects in which their meanings are related. (Recall Sweetser’s idea that metaphorical extension requires some kind of ‘common structure’ across different domains.)

Both of Kratzer’s arguments are therefore ineffective against our favoured polysemy account of modals. Her first argument applies only to views positing a much greater extent of ambiguity than our account. And her second argument applies only to views positing homonymous modals, and not to polysemy views. It would be interesting to investigate how effective both arguments are against ambiguity accounts of modals other than that discussed here, but limits in space urge us to move on to the third argument against ambiguity.

4.3 The cross-linguistic argument

The third argument against ambiguity is based on cross-linguistic data and has been stated most clearly by Valentine Hacquard (2011). It is based on the claim that many languages contain polyfunctional modals:

Do English speakers store two different mays in their lexicon (either as homonyms or polysemes)? […] [Polyfunctionality] is common enough cross-linguistically, and in languages from different families, so as to make a lexical ambiguity account unlikely: it is highly improbable that the same lexical accident should be found in language after language. Rather, it seems that we should give a single meaning for those modals that show an ambiguity, and derive the variety of flavors via some contextual factors […]. (2011: 1489)

This argument can be schematised as follows:

(P1) Many languages from different families feature polyfunctional modals.

(P2) If such modals were ambiguous, widespread polyfunctionality would be an improbable (and inexplicable) lexical accident.

(C) So we should hold that these modals are univocal.27

We will now show that both premises of this argument should be rejected. First, we will present typological data that do not fit with the claim that polyfunctionality is common in languages from different families. Then we will argue against (P2): even if (P1) were true, this would be compatible with a view on which modals are polysemous.

4.3.1 Cross-linguistic data

According to Hacquard, polyfunctionality is “quite pervasive across languages” (2011: 1486) and common in languages from different families. In support of this claim, Hacquard lists examples of polyfunctional modals in French, Italian, Malay, Cairene Arabic and Tamil and cites a number of linguistic studies (Fleischman 1982, Traugott 1988, Bybee et al. 1994 and Palmer 2001). However, the most recent and extensive cross-linguistic study, due to van der Auwera & Ammann (2011), points in a different direction. This study surveys 207 languages around the world and finds only 36 languages containing polyfunctional modals for possibility and necessity, 66 languages containing polyfunctional modals for either possibility or necessity (but not both), and 105 languages containing no polyfunctional modals at all.28

Fig. 3: Map 76A from van der Auwera & Ammann 2011.

How should these results be interpreted? The study finds fewer languages with some polyfunctional modals (102) than languages without any (105), so the claim that polyfunctionality is common or even pervasive across languages seems hard to sustain. Such doubts are strengthened if one takes into account that polyfunctionality is such syntactic properties of modals were taken to indicate different meanings, even fewer languages would be classified as having polyfunctional modals. Second, the study distinguishes only two types of modality: epistemic modality and situational modality, which includes both deontic and dynamic modality. There may thus be languages with polyfunctional modals for different types of situational modality that the study classifies as non-polyfunctional. This latter restriction can be ignored in the current context, as it seems that only languages featuring fully polyfunctional modals (i.e. modals that can express the same kinds of modality as English modals, including epistemic and deontic modality) seem suited to be used in Hacquard’s argument.

27. Arguments of this kind have recently also been employed by Papafragou (1998: 371), Schaffer (2011: 208) and Chrisman (2012: 436).

28. Two things are worth noting about van der Auwera & Ammann’s study. First, their classification disregards syntactic differences between different occurrences of an expression: a modal is classified as polyfunctional even if the different flavours of modality it can take are tied to distinct syntactical constructions (as is the case with the Modern Greek ‘boró’, which translates to ‘can’). If
absent from either possibility or necessity modals in the 66 languages in the middle group. The study does not indicate whether these languages lack modals in one of the two classes (possibility or necessity) completely, or whether they contain modals in both classes, and contain polyfunctional modals in only one of those classes. But at least some languages are of the second kind, containing only non-polyfunctional necessity-modals or only non-polyfunctional possibility-modals.\(^{29}\) This means that there are many more languages with some non-polyfunctional modals than there are languages with some polyfunctional modals (among those studied).

A second suggestive result of the study has to do with the geographical distribution of languages with polyfunctional modals. Polyfunctionality is common in Indo-European languages, but it is rather uncommon in languages from many other families. For example, it is almost entirely missing from Native American languages (see Fig. 3). So polyfunctionality is in fact rather uncommon in languages from different families. This means that out of the examples of polyfunctional modals mentioned by Hacquard and discussed in the literature, only those of Indo-European languages are representative, while polyfunctional modals in languages of other families should be seen as outliers.

We do not want to overstate the significance of this study. After all, even a sample size of 207 languages is fairly small if one takes into account that there are more than 7,000 living languages in total. But it should be clear that the burden of evidence is on the side of those using polyfunctionality to argue against ambiguity. As it stands, the most extensive typological study of modal language (as far as we know) offers more reasons to deny widespread polyfunctionality in languages of different families than to accept it. We thus conclude that (P1) should be rejected.

\(^{29}\) Two examples in this class are Hungarian, which contains only non-polyfunctional necessity-modals, and Ainu, which contains only non-polyfunctional possibility-modals. See van der Auwera & Ammann (2011).

\(^{30}\) Amaral (2008) gives a similar response to a similar argument against the ambiguity of ‘the’. 

4.3.2 Polysemy and cross-linguistic polyfunctionality
Premise 2 said that cross-linguistic polyfunctionality is inexplicable on any views that fall short of univocality. For the sake of disputing this premise, we shall temporarily accept the claim of cross-linguistic polyfunctionality. The gist of our argument is that univocality is not the only plausible explanation for widespread polyfunctionality; polysemy does just as well.\(^{30}\) To explain polyfunctionality, we need to show that the different meanings of modals are somehow or other related, in ways that will impress speakers of different languages and language families. We should not be surprised, either, if the polysemy of ‘healthy’ or ‘long’ is cross-linguistically stable. After all, it is natural to group these meanings together: they are related in ways that may well be cross-culturally acknowledged. It will by now not come as a surprise that the same, on our view, can be said of the different modal flavours.

In Section 3, we have presented some ways in which the modal flavours may be related to each other. Each of these ways seems natural enough to occur across different languages. For example, Sweetser’s (1990) account of metaphorical extension works as well for other languages as it does for English. If English root modals were historically suited to cover non-root meanings by metaphorical extension, then the same seems plausible for other languages. Metaphorical extension is not tied to a certain language, as is also suggested by the fact that many other metaphors are used similarly across linguistic boundaries (e.g. ‘foot’ and its cognates, which in many languages can be applied to a body part and, by metaphorical extension, to a part of a mountain; see Amaral 2008: 290f). In the same way, Traugott’s (1989) explanation of polyfunctionality in terms of pragmatic strengthening, where modal expressions each start out with one modal meaning and gain a second modal meaning after first carrying it as a conversational implication, is not language-dependent.

We submit that the explanations which we have sketched
How many meanings for ‘may’? The case for modal polysemy

5. Philosophical implications of modal polysemy

If our arguments in this paper up to now have been successful, then the polysemy thesis is a serious contender in the semantics of modal language, on at least equal footing with the Kratzerian claim of mere context-sensitivity. We believe that it is indeed, for the reasons outlined in Section 3, the better view. But even the weaker claim that it is a serious contender has some important implications which we would like to draw out in this final section. Section 5.1 addresses implications for the philosophy of language (beyond the obvious modifications that might be required in a Kratzerian framework); Section 5.2 looks at ramifications for the metaphysics and epistemology of modality.

5.1 Modal polysemy and the philosophy of language

In the philosophy of language, modal polysemy allows non-standard accounts of modals to respond to certain objections. First, there is truth-relativism, which has been defended by Andy Egan (2007), John MacFarlane (2011) and others in the on-going debate about epistemic modals. Consider the following variation of our sentence (1):

(7) Holmes might travel to Paris.

Truth-relativists hold that the semantic value of (7) is invariant across contexts and that it is a non-classical proposition whose truth-value can shift with the context (of use or of assessment). The proposition expressed by (7) is true in a context of assessment in which it is compatible with what is known that Holmes travels to Paris, and it is false in a context of assessment in which it is incompatible with what is known that Holmes travels to Paris. While the debate about epistemic modals has focussed on ‘might’, truth-relativists should, it seems, analyse epistemic uses of ‘may’, ‘must’ and other modals in a similar manner.

In a recent paper, Jonathan Schaffer (2011: 206–209) has used the alleged fact that modals are univocal to argue against truth-relativist accounts of epistemic modals. Schaffer’s argument relies on the assumption that truth-relativists treat epistemic modals differently than modals of other flavours: they posit “special semantical clauses” (208) for epistemic modals, which they do not posit for deontic or dynamic modals. For example, they provide different semantics for epistemic uses of ‘may’ than they do for deontic uses. But, Schaffer argues, because modals are univocal, “treatments requiring a special ‘epistemic modal’ element may be eliminated from the start” (208).

Our result of modal polysemy means that truth-relativists can give a quick and easy response to this kind of argument: Because modals are polysemous, we should not be surprised to find different semantic treatments for modals of different flavours. Of course this is not the only way out for truth-relativists. A second response would be to extend the relativist semantic analysis to cover modals of all flavours. MacFarlane (2014) seems to be moving in this direction by providing a relativist semantics for deontic ‘ought’. But not all truth-relativists will want to make this move, and for those who do not the polysemy reply seems perfect.

Schaffer’s argument is aimed not merely at truth-relativism, but at any semantic theory that treats epistemic modals differently than modals of other flavours. These views include the (non-Kratzerian) contextualist account of Keith DeRose (1991), which takes epistemic modals to have special indexical characters, and expressivist accounts (defended e.g. by Hare 1967 and Schnieder 2010), which entail that epistemic modals express an epistemic stance the speaker has towards the proposition expressed by the sentence she utters, rather than

### 31. Schaffer is concerned with ‘might’ and argues that ‘might’ can be used to express epistemic, deontic and dynamic modality. We are not so sure that ‘might’ can be used in these ways and thus prefer to stick to ‘may’ as an example. This switch should make no difference to our discussion of Schaffer’s argument.
How many meanings for ‘may’? The case for modal polysemy

contribute a semantic value to that proposition. And the potential scope of the argument extends even further. Schaffer is concerned only with special treatments of epistemic modals, but it seems like this kind of argument could be brought forward against any account treating modals of one flavour differently than modals of another. So if it turned out that modals were univocal and the argument were successful, this would affect theories in many areas of semantics. However, we hope to have provided strong reasons against the hypothesis of modal univocality, and thus also strong reasons to resist an argument of this kind.

There are further theories in the philosophy of language that benefit from modal polysemy, namely theories of pragmatic invariantism. Such theories, which have been defended by David Braun (2012, 2013) and Kent Bach (2011), entail that the semantic values of sentences such as (1) are invariant across contexts (like relativist approaches), but they also entail that speakers can use such sentences to assert or communicate distinct traditional propositions with absolute truth-values (unlike relativist approaches). One of the propositions a speaker might communicate by uttering (1) is given by (8):

(1) Holmes may travel to Paris.
(8) For all Moriarty knows, Holmes may travel to Paris.

Pragmatic invariantists thus agree with contextualists about the flexibility of modals, but hold that it should be given a pragmatic rather than a semantic explanation.

How does modal polysemy help theories of pragmatic invariantism? First, many theorists will disagree with the claim that a modal such as ‘may’ has the same semantic value on any use. But pragmatic invariantists have to accept fully invariant semantic values only if they take modals to be univocal. A less radical view sees modals as polysemous and merely each meaning of a modal (corresponding to a modal flavour) as semantically invariant. Some theorists might be less concerned about such a less radical view than they would be about a theory that takes modals to be fully invariant.

Second, views of pragmatic invariantism are often criticised for being unable to provide a systematic connection between the semantic value of a sentence and the proposition a speaker can assert or communicate by using it. If they accept modal univocality, pragmatic invariantists have to say that ‘may’ and other modals each have a single semantic value, which speakers can use to assert propositions of any modal flavour. The connection between the semantic value of a modal sentence and the proposition asserted is therefore indeed very loose. Once polysemy is taken on board, this connection can be tightened: if a speaker, speaking literally, utters a modal sentence, the same modal flavour will feature in the semantic value of the modal and the proposition asserted. This is a second way in which modal polysemy can add to the attractiveness of pragmatic invariantism.

5.2 Modal polysemy and metaphysical modality

A central topic in modal metaphysics concerns the relation between metaphysical possibility and necessity, on the one hand, and conceivability and a priori knowability, on the other. There is, of course, wide agreement that these are distinct phenomena: as Saul Kripke (1980) has shown, metaphysical necessity does not entail a priori knowability, nor does conceivability (in a natural sense of the word) entail metaphysical possibility. But the right way of understanding the relation between them is still a matter of controversy.

One area of controversy concerns the metaphysics of the phenomena: Are possibility and conceivability just differentiations of one and the same overall phenomenon, modality? This line, sometimes

---

32. An example is a dispositional semantics for dynamic modals, as outlined in Vetter (2013, 2015).
33. Bach accepts modal univocality; he states that “the fact that there are various sorts of possibility does not show that ‘possibility’ — or ‘might’ or ‘possible’ — has various meanings” (2011: 31). Braun (2012: 462) appears to favour a univocal view of ‘might’, but Braun (2013) mentions the possibility of a polysemous view (without arguing for it).
referred to as modal monism (Goff & Papineau 2014), is taken by two-dimensional semanticists such as Frank Jackson (1998) and David Chalmers (2010). Or are they distinct and mutually independent phenomena, as Dorothy Edgington (2004) and Goff & Papineau (2014) have argued?

A second, closely related question concerns the role of conceivability in the epistemology of metaphysical modality: Conceivability does not straightforwardly entail metaphysical possibility, but can it still serve as a guide to metaphysical possibility when suitably qualified? Modal rationalists, such as Jackson and Chalmers, hold that it does. Anti-rationalists, such as Timothy Williamson (2007) or Sonia Roca-Royes (2011), hold that it does not. The issue has direct impact on debates about the metaphysics of modality. Philosophers who hold, for instance, that the laws of nature are metaphysically necessary (Shoemaker 1998, Edgington 2004 and Bird 2007, to name a few) are committed to anti-rationalism, since it seems clearly conceivable (whatever our restrictions) that the laws of nature had been different. Similar arguments can be made for various alleged de re necessities such as the necessity of origin (see Roca-Royes 2011).

How does the polysemy thesis relate to this debate? We can begin by noting that there is an intimate connection between the dynamic modals and metaphysical modality, on the one hand, and between epistemic modals and conceivability/a priori knowability, on the other. Indeed, authors in the debates just sketched often refer to the epistemic phenomena at issue as ‘epistemic modality’. The distinction between epistemic modality in this sense and metaphysical modality is often introduced with the help of the locutions ‘what might be’ (epistemic, in the linguist’s sense) and ‘what might have been’ (dynamic, cf. Stalnaker 1981: 99). Edgington specifies in more detail how we obtain an understanding of metaphysical modality from a pre-theoretical understanding of dynamic modals:

The [relevant] pre-philosophical notion is that certain things can happen, other things can’t; people and other objects can do certain things and can’t do others. [...] This car can do a hundred miles per hour (though it never will), the other car can’t—as they are presently constituted. Later, when the first has deteriorated and the second hot- ted up, the position may be reversed. [...] Call something absolutely metaphysically impossible if it is metaphysically impossible [in the sense of the above examples] at all times. Its negation is absolutely metaphysically necessary. (Edgington 2004: 6)

In general, we can think of metaphysical modality as a generalization of dynamic modality, abstracting at least from times (as Edgington points out) but also from the restrictions of an individual context (we would like to add). Analogously, we can think of conceivability and a priori knowability — let us call it broadly epistemic modality — as a generalization of epistemic modality in the linguist’s sense — or narrow epistemic modality — abstracting, presumably, from a subject’s particular state of information.34

Given this close connection between the linguistic and the metaphysical distinctions, the polysemy thesis should be a welcome ally to those who reject modal monism and modal rationalism. It can be an ally to them in three different ways. First, it provides a natural motivation for a non-monist conception of modality. Second, it helps the non-monist and the anti-rationalist explain away appearances that may seem to point towards monism. Third, it helps defuse a potential objection. We will briefly sketch each of these three kinds of considerations.

34. In this paper, we have been concerned only with modal auxiliaries and semi-auxiliaries. Of course, in the debate about metaphysical modality, other kinds of modal expressions also play a role: most notably, the adjectives ‘possible’ and ‘necessary’, and their corresponding adverbs. We expect that very much the same kind of argument that we have given in this paper will apply to those expressions as well. Our criteria (1)–(3) from Section 3 obviously apply here as well, though the application of criteria (4) and (5) would require further inquiry.
First, motivation. The polysemy thesis posits a kind of dualism in linguistic modality, which can be seen as an indicator of a corresponding dualism in the metaphysical realm. After all, if we start with two distinct phenomena—dynamic and (narrow) epistemic modality—and generalize from them, there is no particular reason to expect that in so generalizing we will reach broader phenomena that are interlinked in the ways required by modal monism and modal rationalism.

Second, explanation. The polysemy thesis offers the modal dualist and the anti-rationalist an explanation of why these broader phenomena nevertheless seem to have some connection. Whatever relation holds between dynamic and epistemic meanings of modals, be it metaphorical extension, pragmatic strengthening or any other form of explanation, may well hold, in the same or similar ways, at the generalized level as well. It is important here that the relations to which we have pointed are cognitive and pragmatic: there is a relation for us, which explains our grouping these phenomena together, but there is no need to assume that this relation is an interesting metaphysical one. The modal anti-rationalist and anti-monist can thus argue that their opponents’ grouping together of metaphysical and broadly epistemic modality is based on a mistaken, though understandable, failure to properly distinguish the linguistic phenomena referred to by the introductory locutions: what might be, and what might have been.

While it provides some initial motivation and support for a dualist and anti-rationalist view, we hasten to point out that the polysemy thesis does not decide the debate between modal monists/rationalists and their adversaries; we take it that both parties can accommodate polysemy. The case is obvious for the modal dualist and anti-rationalist, who may, of course, think of dynamic and epistemic modality as latching on to two very different kinds of phenomena. However, for the modal monist, two-dimensional semantics provides a crucial distinction between quantifying over possible worlds “considered as actual”, yielding broadly epistemic possibility and necessity; and quantifying over them “considered as counterfactual”, yielding metaphysical modality. While two-dimensionalists have not generally been concerned with the semantics of ordinary-language modals, their apparatus easily lends itself to capturing the polysemy of ordinary modal language.35

Modal polysemy, then, does not entail modal dualism and/or anti-rationalism. Rather, the problem runs the other way around. It is the assumption of univocality which is difficult to accommodate for the anti-monists and anti-rationalists, at least if they continue to think of metaphysical and broadly epistemic modality, along the lines just suggested, as generalizations of dynamic and narrowly epistemic modality respectively. For if modals were univocal, then each would have just one meaning (i.e. character) of which the dynamic and the epistemic readings would be but different contextual specifications: both might be, for instance, a matter of quantifying over possible worlds, and differ only in the restriction on the relevant possible worlds, or (on two-dimensionalist semantics) the angle that is taken on them. It would then be hard to escape the conclusion that they are, at the most general level, just aspects of a shared phenomenon: modality, as captured (on this view) by the whole range of possible worlds. In generalizing from either dynamic or narrowly epistemic modality, we may reasonably expect to reach that very same phenomenon, thus confirming modal monism.

Given the polysemy thesis, however, it is possible to think of the truth-conditions for dynamic and narrowly epistemic modals as coming radically apart, so that generalizing the one is likely to lead to very different results from generalizing the other. Suppose, for illustration, that a dispositionalist account of dynamic modals and an expressivist account of epistemic modals are correct; there is then certainly no reason to believe that what we express with dynamic and epistemic modals are just different specifications of one and the same

35. For the modal monist, the distinction in two-dimensional semantics between quantifying over possible worlds “considered as actual” and quantifying over them “considered as counterfactual” may be construed as capturing the polysemy of modal language. The opponent of modal monism will, however, be open to more radical forms of polysemy, where we do not even construe all modal flavours in terms of quantification over possible worlds.
phenomenon, and that in generalizing from both of them we will reach the same, or even closely related, phenomena.36

Our claim, then, is not that the polysemy thesis definitively establishes the truth of any of the rival views in the debate about the metaphysics and epistemology of modality. But it does open up space for a much more heterogeneous picture of modality than that adopted by modal monists and, typically, by modal rationalists. It is, therefore, a natural ally for those who, like Edgington, believe that metaphysical and broadly epistemic modality come apart, and who think of these two kinds of modality as generalizations of the linguistic categories of dynamic and narrow epistemic modality.

Conclusion
We have argued that modals such as ‘can’ and ‘may’ are polysemous: they have distinct though systematically related meanings, which correspond to the dynamic, deontic and epistemic flavours of modality (and perhaps others in addition to them). Modals meet our criteria for polysemy, and the standard arguments against a polysemy conception can be rejected.

Much work remains to be done, of course. The polysemy thesis raises a number of important questions: Exactly how are the different meanings of modals connected? And how is it that, on a given occasion, one meaning rather than another is selected? These and other questions can be asked for any polysemous expression, and our understanding of polysemous modals should benefit from a better understanding of polysemy in general.

We hope that a further general lesson may be drawn from this paper: Not all ambiguities are alike, and the much-used example of ‘bank’ is an unhelpful paradigm in debates about whether or not a given expression is ambiguous. In claiming that modals are ambiguous, we are not aligning them with the relatively accidental types of ambiguity exemplified by ‘bank’. The ambiguity of modals is of a more interesting and systematic kind, and we expect most other philosophically interesting ambiguities will be of that kind as well. Ambiguity theses cannot be simply dismissed by arguing that an expression does not behave in the way that ‘bank’ does.37

References
Braun, D. 2013: Invariantism about ‘can’ and ‘may’ (as well as ‘might’). Linguistics and Philosophy, 36, 181–185.

37. For helpful discussions of the material in this paper, we would like to thank audiences at Dortmund, Norwich and Berlin, as well as Mathias Böhm, Cian Dorr, Angelika Kratzer, Christian Nimtz, Wolfgang Schwarz, Nathan Wildman, two anonymous reviewers for this journal, and the participants in the research colloquia of Olaf Müller, Barbara Vetter and Emanuel Viebahn at the Humboldt University of Berlin. Both authors contributed equally to this paper.
How many meanings for ‘may’? The case for modal polysemy


How many meanings for ‘may’? The case for modal polysemy
