21
Modality
Ilse Depraetere

1 Introduction

Modality is among the most widely studied topics in linguistics. It is a concept that is used to designate a wide array of language facts. In his overview of the field, Portner (2009: 2–8) mentions the following examples that involve modality and groups them in three classes:

- modal phenomena at sentence level (modal auxiliaries such as *may*, *must*, *can*, *will*, *should*; modal adverbs like *possibly*, *probably*; generics, habituals; tense and aspect (the future, the use of the past to express unreality, as in *Even if you stayed until tomorrow, I’d be sad*, the progressive, the perfect); conditionals; covert modality (e.g. *Tim knows how to solve a problem* contains an implicit expression of ability, as is clear from the paraphrase *Tim knows how he can solve a problem*).
- modal phenomena at sub-sentential level (modal adjectives and nouns, propositional attitude verbs (e.g. *believe*, *hope*), and adjectives (e.g. *certain*, *pleased*), verbal mood (e.g. the subjunctive), dependent modals, as in *I’d be surprised if David should win*, negative polarity items).
- discourse modality (evidentiality (cf. section 2.2), specific functions of clause types).

The diverse range of potentially modal phenomena is also clear from research on modality from a cross-linguistic perspective (e.g. Bybee et al. (1994), Palmer (2001), de Haan (2005)).

It is not only the breadth of the field that is striking: modality is a phenomenon that has inspired research in widely divergent theoretical frameworks, and in neighbouring disciplines such as logic and philosophy. Modality has been studied extensively in:

- Formal semantics (Kratzer 1991, 2012)
- Cognitive linguistics and force dynamics (Talmy 1988; Langacker 2003)
- Pragmatic theory (Papafragou 2000)
In other words, both the phenomena that are examined under the heading of “modality” and the theories that have served as a framework for analysis are multiple and varied. Given the breadth of the field, a functional or semantic definition of modality encompassing all its potential formal realizations is not easy to achieve (cf. e.g. Narrog (2005), Nuyts (2005), and Declerck (2011) for useful discussion). A common characterization of modal verbs, which are fairly uncontroversial expressions of modality, is that these are forms the meanings of which hinge on the notions of possibility or necessity. However, even when restricted to the core set of English modals, the definition is already problematic: the meanings of shall and will cannot be captured, in a straightforward way, in terms of possibility or necessity (see Chapter 20). A broader definition is that modality refers to forms that represent situations as non-factual. Here again, even when restricting oneself to the core modals in English, a definition along these lines is not without problems: some (e.g. Salkie (2014)) have argued that can communicates enablement and does not express non-factuality at all.

Mood refers to the grammatical marking of specific meanings through affixes attached to the verb: certain affixes express the meaning of non-factuality; others express the meaning of factuality. In the traditional view, there are three moods in English: the indicative mood (e.g. She got her PhD in 1993), the imperative mood (Leave me alone!) and the subjunctive mood (Parliament decreed no more soldiers be sent to the war zone). The mandative subjunctive, illustrated in the example, is more frequent in American English than in British English (cf. Leech et al. (2009)). From a synchronic point of view, there seems little formal justification to recognize an inflectional subjunctive mood in present-day English as the form of the present subjunctive corresponds to the bare infinitive. Be is the only verb that has a so-called past subjunctive, and its form corresponds to the past indicative form were, as in I wish you/she were here. Some have argued that it is more accurate to speak of a subjunctive clause type (Aarts 2012) or subjunctive constructions (Huddleston et al. 2002: 51, 993) rather than subjunctive mood. (The same line of meaning could be developed for the imperative (mood) in English.) Other languages have a wider range of moods, such as the optative in Turkish, to express a wish, a hope or a command. Nenets, a language spoken in Northern Russia, has as many as sixteen moods. Rothstein and Thieroff (2010) offers an overview of morphological mood in 36 European languages. Not all moods express modal meaning. The indicative, for instance, is typically used to represent a situation as a fact, which is clearly not modal. In sentences like If I got the grant, I would be absolutely thrilled though, got loses its unmarked temporal meaning: it no longer refers to the past, but to the future, and it no longer presents a situation as a fact, but as hypothetical. For these reasons, this form is labelled as a “modal indicative”, at first sight a contradiction in terms. While “modal” captures the semantics, “indicative” refers to the morphological make-up – it is standard (but not universal) practice for “mood” to refer to a verbal inflection. (cf. Thieroff (2010)) A similar form-function tension is observed in the chapter on tense: not all instantiations of a particular tense, for instance, the present tense, refer to present time, but that does not stop one from referring to the verb in The performance starts at eight as a present tense form.

The conceptual pair “mood and modality” is similar to “tense and time”, in that in both cases, the first item refers to a formal inflection and the second captures a specific function or meaning: that of expressing time (in the case of tense) and that of expressing (non-)modal meaning (in the case of mood). As will be clear from the overview at the beginning of this section, mood does not exhaust the list of forms that express modal meaning, in the same way as tense is not the only formal means to express time or a temporal relation (cf. Chapter 20).
“Modality/mood” is one of the components in the commonly used acronym TAM, the association with tense and aspect pointing to the interaction between three phenomena that pertain to the verb (phrase) (cf. e.g. Hogeweg et al. (2009)), even though one must bear in mind that tense and mood are formal categories while aspect and modality are notional categories. The discussion so far has already touched upon a link between tense/time and modality: specific tenses can be used in contexts that express non-factual meaning. The present tense, like the past tense (cf. example above), can express hypothetical meaning (He behaves as though he knows everything about it – hypothetical). The past tense and the past perfect can communicate counterfactual meaning (If I knew how to replace a flat tyre, I wouldn’t be stuck here. – I wish you had told her.). The debate about whether English has a future tense (see Chapter 20) illustrates another link: here the question is whether, in the absence of an inflection that signals future time in English (or any language that does not have an inflectional morpheme for future time reference), and given the fact that a future situation is by definition not factual, will should be considered as a marker of modality rather than as a tense marker. Because non-present tenses are removed from the present and/or from reality, they have sometimes not been defined in terms of temporal relations, but in terms of detachment. In cognitive linguistics, the present tense has been analysed in terms of epistemic immediacy (cf. e.g. Langacker (2011) for discussion).

As for the link between aspect and modality (see Chapter 19), among the relations that have been studied is that between mood and modality and the imperfective/perfective (cf. e.g. Abraham and Leiss (2008)); the progressive in English has been analysed as a kind of modal operator as well (Dowty 1977).

In this chapter, the focus will be on verbs that express modal meaning in English. (Hansen and de Haan (2009) offer a description of the system of modals and modal constructions in languages spoken in the European area.) Despite extensive research in this subfield of modality, there are still a number of outstanding problematic areas:

- The more theoretically oriented work tends to be tested on a relatively small data sample, and it is not very clear if it can cope with a richer data set.
- Even meticulous empirical studies at times uncritically incorporate concepts such as subjectivity, source of modality or strength of modality into the analysis. While there seems to be unanimity that these concepts are crucial, there is still a need for careful clarification and operationalization for empirical (corpus) analysis.
- A key question in the characterization of modal expressions is the extent to which their meaning is determined by the independent lexical semantics of the verbs and to what extent it is contextually determined. In other words, do modal verbs have one core meaning that is contextually enriched (monosemy) or are the different meanings sufficiently differentiated semantically to consider modal expressions polysemous (see Chapter 13)? Answering this question means taking a theoretical stance towards the semantics/pragmatics interface (see Chapter 10).

The outline of the chapter is as follows. After a brief overview of the formal characteristics of modal verbs, I will present a few of the controversies in research on modal verbs in more detail; that is, questions of taxonomy, the polysemy/monosemy debate and the notions of subjectivity and strength. Section 4 draws a sketch of the historical development of modal meanings and grammaticalization. The conclusion touches upon the question of the semantics/pragmatics interface and gives an indication of possible routes for future research.
2 The meaning of modals: some controversial issues

2.1 A note on the formal behaviour of modals in English

Before examining the semantics of modal verbs, it is useful to sketch their formal profile. The core modals (*can, could, may, might, must, will, wouldn't, should*) do not require do-support in four specific contexts: in negative sentences (*I can't help it*), in contexts in which there is subject-verb inversion (*Can you come?*), in contexts in which the main verb is unexpressed (*Can you come? – Yes, I can*), and in contexts in which a contrast is established (*You can't uncork a bottle with a fork. – Now watch this. See? I can uncork a bottle with a fork*). This is a feature they share with the auxiliaries be and have. They exhibit a further number of formal features:

- they do not inflect for person and number (*She can/*cans sing well);
- they cannot be stacked (except in some varieties of English) (*She must can be available tomorrow);
- they do not have non-finite forms (*to can; *She is canning);
- they are followed by a bare infinitive (*She can sing/* to sing).

*Ought to, need and dare* are sometimes said to be peripheral modals, for different reasons: *ought* does not have all of the prototypical formal features; it is followed by a to-infinitive. The constraint on the latter two auxiliaries is that they only feature in non-assertive contexts; that is, sentences that are negative and/or interrogative (*Need(n't) I explain? – You needn’t explain. – *You need explain*.). There is a further group of expressions with even fewer of the features of the core modals: modal expressions with be, including *be to, be supposed to, be able to, be permitted to, be liable to* are sometimes labelled quasi-modals or semi-modals. *Have to and need to* are lexical verbs that express the modal meaning of necessity; they are sometimes also included in the category of semi-modals. This brief overview does not provide an exhaustive list of verbal expressions that communicate modal meaning, the list including further verbs like *want to, have got to*, etc. (cf. section 3).

2.2 Taxonomies of modal meaning

Some classifications of modal meaning (e.g. Coates (1983)) make a binary distinction between epistemic modality and non-epistemic or root modality. The label “root” modality reflects the historical fact that non-epistemic meaning is the more basic one, epistemic meanings usually having developed out of non-epistemic ones (cf. section 3). Inspired by Palmer (1990), other authors advocate a three-fold classification and distinguish epistemic modality, deontic modality and dynamic modality (cf. e.g. Nuyts (2000), Verstraete (2001), Huddleston et al. (2002), Collins (2009)). An approach along these lines implies that the distinction between deontic and dynamic modality is considered to be as important as that between epistemic and deontic meaning or that between epistemic and dynamic meaning.

<table>
<thead>
<tr>
<th>epistemic possibility/necessity</th>
<th>non-epistemic or root possibility/necessity</th>
</tr>
</thead>
<tbody>
<tr>
<td>epistemic possibility/necessity</td>
<td>deontic possibility/necessity</td>
</tr>
<tr>
<td>dynamic possibility/necessity</td>
<td></td>
</tr>
</tbody>
</table>
Modality is said to be *epistemic* if it communicates how likely or unlikely it is, according to the speaker, that a situation is the case. In other words, the scope of the modality is wide: the proposition is said to be necessarily (He must be older than 50 – It is necessarily the case that he is older than 50), possibly (He may be older than 50 – It is possibly the case that he is older than 50) or impossibly (He can’t be older than 50 – It is impossible that he is older than 50) the case in the speaker’s opinion. In some definitions, the notion of truth is foregrounded: “Epistemics are clausal-scope indicators of a speaker’s commitment to the truth of a proposition” (Bybee and Fleischman 1995: 6). While the delineation of epistemic modality with respect to the non-epistemic meanings of dynamic modality and deontic modality is relatively unproblematic, there is no unanimity on whether so-called evidential modality is a subclass of epistemic meaning or whether it is on a par with it. Nuyts (2000) defines this category of modal meaning as follows:

> Evidentiality concerns the speaker’s indication of the nature (the type and quality) of the evidence invoked for (assuming the existence of) the state of affairs expressed in the utterance. This does not involve any explicit evaluation in terms of the state of affairs being true or not. (2000: 27)

Note that in English, evidentiality is not expressed through modal verbs, unlike in German, for instance (*Er soll krank sein.* – “He is said to be ill.”) but through other markers, as in *Two hundred people are alleged to have died in the plane crash* or *I hear that you’ve been promoted.* In other languages, evidentiality is expressed morphologically (cf. e.g. Willett (1988) for a cross-linguistic survey of markers of evidentiality and Boye (2012) for an in-depth discussion).

*Deontic* modality encompasses permission (on the possibility side) and obligation (on the necessity side). The definitions that have been given of this type of modality often involve reference to speech acts, such as that of giving permission (deontic possibility), or that of giving an order or prohibiting an action (deontic necessity). In other words, for some people, the concept of deontic modality hinges on the pragmatic notion of (indirect) performativity, witness the fact that the meanings can be adequately glossed with a performative verb, *You must be back by twelve* corresponding to *I oblige/require/demand you to be back by twelve* and *You can park your car in front of my garage* corresponding to *I authorize you to park my car in front of my garage* or still *You can’t have more cake* corresponding to *I forbid you to have more cake*.

If it is the pragmatic feature of “directive speech act” that is constitutive of “deontic necessity” and which sets it apart from dynamic necessity (cf. below), then one might wonder if Coates’s (1983) view is not appropriate: she argues that the modal logic term “deontic”, used by von Wright (1951) to analyse the logical relations between obligation, permission or prohibition, is inappropriate to discuss the semantics of (what she calls) root (that is, non-epistemic) modality. As she sees it, cutting up root necessity meanings, for instance, obscures the essential unity of root necessity utterances, which share a common, basic meaning; that is, “it is necessary for”. Moreover, the cut-off points between the subclasses are bound to be arbitrary. A point to be borne in mind is that in some classifications of modal meaning, deontic necessity (or, for that matter, deontic possibility) is a subclass of root necessity (root possibility), while in other taxonomies, it is a cover term that encompasses both “performative” and “non-performative” instances of necessity (possibility). Verstraete (2001) offers a very illuminating discussion, whereby a distinction is drawn between “modal
performativity” and “interactive performativity” (relating to the speech act value) and the role they have to play in the definition of modal meaning (cf. section 2.4).

Dynamic modality is concerned with “properties and dispositions of persons, etc. referred to in the clause, especially by the subject NP” (Huddleston et al. 2002: 178). Ability (She can play the flute) and volition (She will help me with the BBQ) illustrate the “clearest cases” on the possibility side, but dynamic possibility also includes “what is reasonable” (You can always say you’re too busy), “what is circumstantially possible” (Water can still get in) and “what is sometimes the case: the ‘existential’ use” (These animals can be dangerous). A “clear” example of dynamic necessity is Ed’s a guy who must always be poking his nose into other people’s business; this category also includes “circumstantial necessity” (Now that she’s lost her job she must live extremely frugally) (Huddleston et al. 2002: 184–5). All authors who distinguish between dynamic and deontic modality point out that the boundaries are somewhat fuzzy. Gisborne (2007) brings up some fundamental problems with the notion of dynamic modality and Depraetere and Reed (2011) are likewise critical of this category and try to offer a more explicit definition of modal non-epistemic possibility meaning that is “not ability” and “not permission”. To do so, they use three criteria (scope, source, potential barrier), which enable them to pin down five subclasses of root possibility meaning: ability, opportunity, permission, general situation possibility (GSP) and permissibility. The scope is wide or narrow depending on whether the possibility bears on the entire proposition (as in Cracks can appear overnight) or on the VP (as in This printer can also scan documents). “Source” refers to the origin of the possibility, which may be subject-internal, in which case the possibility originates in innate capacities or acquired skills of the subject referent. Subject-external sources may be of different types: an external authority, rules and regulations or circumstances may be at the origin of the possibility. If the source has source status because it can potentially impose a barrier on actualization, it carries the feature [+ potential barrier]. For instance, in All vehicles rented in Ireland may only be driven in Northern Ireland and the Republic of Ireland (situation permissibility), the rental policy regulations owe their status of source to the fact they determine whether there can be actualization or not of the situation. The combination of features that define the different subtypes of non-epistemic meaning are summarized in Table 21.2:

While the triad “epistemic – dynamic – deontic” is a standard classification of modal meaning as expressed by modals, alternative proposals have been put forward. Bybee and Fleischman (1995) and Bybee et al. (1994) distinguish between speaker-oriented modality, which is expressed by forms that mark directives, such as the imperative form in English, epistemic modality and agent-oriented modality, a cover term encompassing what Huddleston et al. (2002) call dynamic modality and deontic modality. There is a further class of “subordinating moods”, which they reserve for forms expressing the speaker-oriented and epistemic modalities used in subordinate clauses, such as complement clauses, concessives

<table>
<thead>
<tr>
<th>Source</th>
<th>Scope</th>
<th>Potential Barrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>external</td>
<td>wide</td>
<td>+ potential barrier</td>
</tr>
<tr>
<td>internal</td>
<td>narrow</td>
<td>– potential barrier</td>
</tr>
</tbody>
</table>

### Table 21.2 Classification of root possibility meanings based on Depraetere and Reed (2011)

<table>
<thead>
<tr>
<th></th>
<th>situation permissibility</th>
<th>GSP</th>
<th>permission</th>
<th>opportunity</th>
<th>ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>external</td>
<td>external</td>
<td>external</td>
<td>external</td>
<td>internal</td>
</tr>
<tr>
<td>Scope</td>
<td>wide</td>
<td>wide</td>
<td>narrow</td>
<td>narrow</td>
<td>narrow</td>
</tr>
<tr>
<td>Potential Barrier</td>
<td>+ potential barrier</td>
<td>– potential barrier</td>
<td>+ potential barrier</td>
<td>– potential barrier</td>
<td>– potential barrier</td>
</tr>
</tbody>
</table>
and purpose clauses. Narrog (2012) identifies nine different subclasses of modal meaning (epistemic, deontic, teleological, preferential, boulomaic, participant-internal, circumstantial, existential and evidential), and they are characterized in terms of two dimensions: that of volitivity and that of speech act orientation.

Observations to the effect that the borderlines between categories are fuzzy, irrespective of the taxonomic approach taken, abound and this brings up a general taxonomic question, whether, ultimately, the gradience is of the “subsective” or “intersective” type (cf. Aarts (2007)): that is, does it relate to degree of typicality (within a class) or to degree of membership? Applied to modal meanings, the question is whether the indeterminate examples constitute less prototypical examples of a specific meaning or whether they are indicative of the existence of an intermediate class.

2.3 Polysemy/monosemy

The examples with can mentioned in section 2.2 show that one modal can communicate a variety of meanings. Many of the English modals are “multi-functional” and this is a phenomenon that has been observed cross-linguistically, but it is not universal and it is not spread evenly. In a typological study of 241 languages, Van der Auwera et al. (2005) have found that 49 languages have forms (including but not restricted to verbs) that are polyfunctional; that is, they can express epistemic as well as non-epistemic meaning, both for possibility and necessity. There are 30 languages that have forms that can express both types of necessity (for four of them the evidence is not conclusive yet as to whether the polyfunctionality is partial or also extends to possibility) and another 39 that have forms that express both types of possibility (for 11 of them the evidence is not conclusive yet as to whether the polyfunctionality is partial or also extends to necessity). One hundred and twenty-three languages do not have polyfunctional forms. It is interesting to observe that in certain languages, one form may be used both for possibility and necessity: for instance, in Danish, må can express both obligation and permission. The following examples from the British National Corpus (BNC) illustrate some of the meaning distinctions communicated by must:

(1) “What an unusual boy he must be,” my mother remarked, “if he has the sense to see beneath the exterior to the person inside.” (BNC) (epistemic)
(2) State House already houses a recruiting office for the Royal Navy and Marines, and visitors must apply for a temporary pass at the security ground floor entrance. (BNC) (non-epistemic)
(3) I feel I must write and thank my fellow citizens for the magnificent turnout at the main Battle of the Atlantic march-past. (BNC) (non-epistemic)
(4) “You must come out of the sun, Mr Gray,” said Lord Henry. (BNC) (non-epistemic)
(5) Whatever the risk, she must go to him and tell him to go back. (BNC) (non-epistemic)
(6) Then she took a quick look at Vern and said to me, “You must come and see my photos, Bina.” (BNC) (non-epistemic)

While must communicates non-epistemic necessity in a straightforward manner in (2) to (5), it is likely that Bina, in example (6), will take the sentence as an invitation rather than an order. In (2), there is reference to a general regulation which is represented as unlimited in time; in (3) the necessity is self-imposed and in (4) and (5) the source of the necessity is
external to the “obligee”. If it is true that the borderlines between modal categories are fuzzy (cf. section 2.2), then it will logically follow that the meaning of modals is fuzzy. Can we say that all modals have one meaning, the instantiations of which are just contextual realizations of the basic semantic core? Or should one argue anyway that modals are polysemous, and that the different meanings which they communicate (at least the distinction between epistemic and non-epistemic meaning) are sufficiently distinct to justify the conclusion that they are autonomous, semantic classes of their own (see Chapter 13)? Here again, opinions differ. The following are among the arguments used by “polysemists”:

(a) Ambiguous examples exist:

(7) The second part must contain the item name. ((a) epistemic reading: “if it’s not in the first part, then the only possible conclusion is that it can be found in the second, as the instructions say that the item name must be mentioned in the description”; (b) non-epistemic reading: “this is the requirement”)

(8) You know there is a tendency for malignant change anyway, and one of the reasons for removing a benign parotid tumour is that it may go malignant, in twenty years’ time or so. (British English component of the International Corpus of English (ICE-GB): S1B-010) ((a) epistemic reading: “(on the basis of information that is the speaker’s disposal, (s)he concludes that) it is possible that a (specific) benign parotid tumor will go malignant”, (b) non-epistemic reading: “the situation of a benign parotid tumor going malignant is conceivable/theoretically possible”)

(b) Scope of negation testifies to the difference in meaning between epistemic and non-epistemic meaning in the following examples:

(9) 7.1 The Agreement shall be binding upon all the parties and may not be assigned without the prior written consent of the other party. (BNC) (non-epistemic: the negation bears on the modal meaning, cf. “it is not permitted/possible”)

(10) These may not be recognized as symptoms of stress – that is, until the stress is reduced and the symptoms disappear. (BNC) (epistemic: the negation bears on the proposition, cf. “it is possible that they won’t be recognized as symptoms of stress”)

Note that this test does not yield the same result with all modals. In both the epistemic and the non-epistemic example with must below, negation bears on the proposition/VP:

(11) She had hinted darkly that Wilson herself must be to blame, that she must have been weak, must not have written plain enough and as she had been instructed (BNC) (epistemic, “it is necessarily the case that she did not write plain enough”)

(12) He would be lost in the city on his own and besides he must not forget about his grandparents—they would die if he did not look after them. (BNC) (non-epistemic, “it is necessary for him not to forget”)

(c) Epistemic and non-epistemic meanings have different paraphrases, which shows that they are semantically distinct (cf. e.g. Coates (1983)). Epistemic meanings are brought out by a paraphrase with a that-clause: it is necessarily/possibly the case that X; the meaning of non-epistemic readings is adequately captured by a for-clause: it is necessary/possible for X to Y.

Polysemists often point out that there are fuzzy borderlines between the modal categories (deontic/dynamic, dynamic/epistemic) (cf. e.g. Huddleston et al. (2002): 179; Collins (2009): 23)
and even though they do not always explicitly say so, the care with which they do seems to sug-
gest that fuzziness as such may somewhat jeopardize the position taken. Monosemists (cf. e.g.
Groefsema (1995: 55–7)) indeed use indeterminacy of meaning as an argument in favour of mon-
osemy, but they also tend to qualify their approach. Papafragou, for instance, writes:

My aim in this chapter has been to strike a middle way between polysemy-based and
radical monosemy accounts of English modals. I have tried to offer a semantics rich
enough to allow for difference in content in various modals, and yet underspecified to
the extent of drawing on extensive pragmatic inferencing until it yields a complete truth-
evaluable representation.

(Papafragou 2000: 84)

To formulate a clear answer to the question, a number of issues need to be resolved: first, it is
necessary to pin down criteria that will define, in an unambiguous way, the modal meanings
involved. As will be clear from the discussion in section 2.2, this is not necessarily an item
that can be ticked off the list. Second, it is necessary to turn to lexical semantics and look for
adequate tests that can identify whether the different meanings taken stock of are indicative
of polysemy or vagueness (see the discussion in Chapter 13). For instance, to what extent
can identity-of-sense tests (cf. e.g. Riemer (2005): 140–1) be used to identify the status of the
meaning distinctions? Third, one needs to settle on the impact of potentially indeterminate
examples: do they fade away once the defining criteria have been fine-tuned and made fully
operational for empirical analysis? If so, this may well pave the way for a polysemic account.
If not, while fuzziness may, at first sight, be less of a problem for a monosemous approach, if
the framework is to be explanatorily adequate, one does need to account for the meaning dis-
tinctions that arise in context anyway, which means that, eventually, monosemists basically
face the same challenge as the polysemists. Fuzziness may well point to historical change,
whereby indeterminate examples illustrate “bridging contexts” (Enfield 2003: 28); that is,
they point to intermediate stages in the historical development of meanings.

2.4 Subjectivity

While subjectivity is a linguistic feature that is pervasive, occurring when a speaker “in making
an utterance, simultaneously comments upon that utterance and expresses his attitude to what
he is saying” (Lyons 1977: 739), it plays a major role in the characterization of modal meaning.

It is very closely associated with epistemic modality, since epistemic meaning reflects the
speaker’s subjective assessment of whether the proposition expressed is the case. The paraphrases
of epistemic sentences bring out the subjective component relatively straightforwardly: That must be Jennifer can be glossed as I confidently conclude (on the basis of situational and
world knowledge to which I have access) that the person entering the house is Jennifer, a para-
phrase of You may be right being I conclude (on the basis of situational and world knowledge
that is at my disposal) that it is possible that your opinion is correct. The subjectivity resides
in the fact that the conclusion formulated is based on a personal assessment of evidence that is
thought to be sufficient by the speaker to warrant the conclusion drawn. Even though subjectiv-
ity is characteristic of most epistemic sentences, examples of objective epistemic modality have
also been signalled, even though they are rare (Coates 1983: 20, 55):

(13) The simple truth is that if you’re going to boil eggs communally, they must be hard.
(Coates 1983: 18)
In this sentence, it is not the speaker’s subjective assessment of a situation that is communicated, but rather what the speaker regards as an empirically verifiable objective fact. (Others have called this intersubjectivity: cf. e.g. Nuyts (2012).) While it is still the speaker who makes an inference, and explicitly presents it like an inference, given the state of the world, this is in fact the only conclusion that can be drawn. Lyons uses the following example to explain objective epistemic modality (1977: 791, 798):

(14) Alfred must be unmarried.

This sentence may be uttered in a context in which this conclusion is the only one that is logically possible. For instance, if Alfred wears no wedding ring and if he is a member of a community in which married people wear wedding rings, then the epistemic judgement is objective. Examples of this type should still be distinguished from alethic modality (cf. e.g. Lyons (1977: 797ff), Hoye (1997: 48–53)), a term used by logicians to refer to the modality of the necessary truth of propositions of the type given in (15) and (16):

(15) If x is a prime number between 90 and 100 it must be 97. (Huddleston et al. 2002: 173)
(16) Alfred is a bachelor, so he must be unmarried. (Hoye 1997: 49)

Subjectivity is not restricted to epistemic modality. On the non-epistemic side, subjectivity is most commonly associated with deontic meanings. When the source of the obligation or the permission is “internal”, in the sense that the obligation imposed or the permission granted or refused reflects an autonomous decision by the speaker, the obligation (You must be back by 10) and permission (You may come in) is subjective. When obligation or permission is seen as coming from some source external to the speaker, it is said to be objective (You have to be back by 11 (these are the house rules) – (pointing to a traffic sign) You can park your car on this side of the street).

The criterion of subjectivity/objectivity is often appealed to in order to differentiate the semantic profile of modal verbs that communicate the same meaning. For instance, have to is said to typically express objective necessity while must is typically used to express subjective necessity; ought to is said to be more objective than should:

(17) You even have to pay extra if you want to have bread with your meal. (ICE-GB, W1B-002) (objective)
(18) You must believe me. (ICE-GB, W2F-008) (subjective)
(19) Appeals Procedure: Appellants ought to have adequate time in which to prepare their appeals, and an effective opportunity to present a counter-argument to the reasons for refusal. (BNC, CFH 166) (objective)
(20) You should stop whingeing and start trying to make things work. (BNC) (subjective)

Similarly, can is associated with “objective permission” examples, with permission may being (usually) subjective:

(21) You can only have showers on week-days after supper, and you have to pay 5 Francs each time. (ICE, W1B-002)
(22) May I sit down for a minute? (ICE, W2F-018)
Such basically introspective intuitions are widespread while clear definitions or indications as to how to operationalize the parameter of subjectivity are often lacking. Verstraete (2001) offers a very insightful discussion and puts forward a test to determine whether a sentence communicates subjective rather than objective modality (2001: 1521–2). When the sentence is turned into a question, in the case of subjective modality, it is the addressee’s commitment to the necessity that is at stake (cf. She must leave the room immediately (subjective) – Must she leave the room immediately? – “Do you want her to leave?”), whereas in the case of objective modality, it is his commitment to the truth of the proposition (cf. Seasnakes must surface to breathe (objective) – Must seasnakes surface to breathe? – “Do you think it is true that it is necessary for seasnakes to surface to breathe?”). Depraetere and Verhulst (2008) and Verhulst et al. (2013) are further initiatives to pin down more explicitly the nature of “objective” and “subjective” modality to arrive at an empirically justified semantic profiling of modal verbs such as ought to, should, be supposed to, have to and must. Narrog (2012: 23–45) also provides very useful discussion of subjectivity in modality. Note that in cognitive grammar “subjectivity” is viewed differently: like tense inflections or person inflections, all modals relate the complement to the “ground” – that is, the speech situation – and are therefore said to be subjective (cf. e.g. Langacker 2003).

2.5 Strength

Strength is also often brought up as a defining criterion in discussions about modal meaning, in two different ways: in the field of modal necessity, some modals are said to be inherently “stronger” than others. No one would disagree with the fact that You should do it is not as forceful or “strong” as You must do it. In other words, there seems to be a scale of strength, and as in the case of subjectivity/objectivity, this feature helps to pin down the semantic profile of the different verbs that express necessity. Identifying the difference in strength between, for instance, must and should, is relatively straightforward and unproblematic. However, the judgement is a lot harder when one draws into the discussion the whole range of verbs that communicate non-epistemic necessity; that is, when we also try to position have to, need to, ought to, be to, be supposed to on the scale of modal strength. In Verhulst et al. (2013: 218), an overview is given of the opposing views on the matter concerning the inherent strength of be supposed to, should and ought to. The authors (2013: 219–21) argue for a definition of strength in terms of the inescapability (or not) of the necessity and the relative gravity of non-compliance. In other words, while it is beyond doubt that necessity modals differ in strength, the route of intuitive appreciation of “strength” is not convincing if this parameter is to function as a reliable and objective criterion to determine meaning distinctions between modals.

Strength also enters the discussion at the level of pragmatic effects: in this case, possibility modals can be strengthened to communicate necessity. For instance, in the context of an oral exam, an instructor saying You can leave now to a student will be understood to be expressing an order (Please leave the room). Likewise, You must come and see my photos, Bina (example (6)), is unlikely to be perceived as expressing a necessity or an order; it will rather be understood as a friendly but firm invitation. In this case, the necessity modal is pragmatically weakened. Note that the status of the strengthened or weakened meaning, while highly conventionalized, is truly pragmatic, in the sense that it constitutes implicated meaning, which, strictly speaking, can be cancelled (even though this is likely to happen in jocular contexts (Can you leave the room? – Yes I can, but I won’t), with the interlocutor (mistakenly) responding to the literal (semantic) meaning.)
3 The diachronic development of modal meaning

With the exception of willan (“intend, want”), the Old English forms corresponding to the core modals can (cynnan, “know” and “have the mental ability, know how to”), may (magan, “have the physical ability to”), must (motan, participant-external “ability” and permission) and shall (sculan, “obligation, necessity, compulsion”) were all preterite-present (lexical) verbs in Old English; that is, they all had a present tense like the past tense of a strong form and a past tense formed on an irregular stem with the endings of the weak past.

(23) present  can  may  must  shall  will
1st/3rd ps sg  can(n)  mæg  mōt  sceal  wille/wile (wille)
2nd ps sg  canst  meaht  mōst  scealt  wilt
plural  cunnan  magon  mōton  sculon (sceo-)  willað
past
1st/ 3rd ps sg  cūðe  meahte or mihte  mōste  sceolde  wolde
2nd ps sg  cūðest  meahtest or mihtest  mōstest  sceoldest  woldest
plural  cūðon  meahton or mihton  mōston  sceoldon  woldon

(Quirk and Wrenn 1989: 55–8)

In the late Middle English period the inflection in the second person singular was lost and the historically past form of must came to be used for present time reference.

It is interesting to observe the so-called drag chain effect in the development of modal meaning in Middle English, summarized as follows by Goossens:

shulen (Modern English shall) gets partially grammaticalized for the future and the conditional, so that moten (the forerunner of must) is taking over the expression of obligation from shulen and clears the way for the use of mowen [ancestor of may] in the field of extra-subjective possibility (including the expression of permission). (…) mowen also retains (and has even expanded) its position in the area of intra-subjective possibility, so that there conne(n) and mowen are largely interchangeable.

(1992: 383)

In other words, during the ME period, the permission uses of must became increasingly restricted and were slowly replaced by may/magan (have the physical ability) and can/cunnan (have the mental ability). Must developed from permission to necessity meaning in the context of negative (denied) permission: there was an inference from “you may not” to “you are obliged not to”. “Invited inferencing”, in the form of quantity implicatures arising from Grice’s second Maxim of Quantity (“Do not make your contribution more informative than is required”) that are conventionalized, have therefore been argued to play a major role in the shift of modal meaning. For instance, if one has permission to do something, this may implicate expectation: “So if I say You may go, I may, in the right circumstances, implicate that I want you to go, and in this sense you have some obligation to go” (Traugott 1989: 51). The frequent collocation of motan with nedes (necessarily) contributed to the development of epistemic necessity meaning and this collocation also led the past form to become used with present time reference. Another factor that influenced the latter shift is that non-epistemic necessity was understood as “coming to be obliged” and epistemic necessity as “coming to believe or conclude” (cf. e.g. Goossens (1982), Traugott (1999), Traugott and Dasher (2002)).
In Old English, *magan* functioned most extensively to communicate ability (internal possibility); that is, physical, perceptual and intellectual capacity, and only marginally expressed permission, some examples combining “internal” and “external” sources (such as circumstances) (Goossens 1987). The “internal possibility” core gradually shrank and the external possibility meaning was extended more and more to include permission (as *motan* was shifting its territory from permission to obligation). As the epistemic function became more important, *may* also developed epistemic meaning.

*Cunnan* only marginally expressed internal possibility (ability, intellectual capacity), its core meaning being “know (a person, a language)”, but it invaded, little by little, the territory of *magan* and came to express possibility that not only depends on certain capacities of the subject, but also on external factors.

Even though it has been pointed out that examples can be found with *magan, sculan* and *willan* that show the first “traces of epistemic use” in Old English, there seems to be a consensus that epistemic meaning develops out of non-epistemic meanings in the fifteenth century, in a process of “subjectification”, which “involves increase in coding of speaker attitude, whether of belief, assessment of the truth, or personal commitment to the assertion” (Traugott 1989: 49). Nordlinger and Traugott (1997) argue that “general participant-external wide-scope root/deontic necessity” should not be overlooked as a phase in the evolution from narrow scope root/deontic meaning to epistemic meaning. In line with this observation, Depraetere and Reed (2011) (cf. section 2.2) distinguish between “wide-scope” non-epistemic possibility, which they label “general situation possibility” and “narrow scope” non-epistemic possibility, which they label “opportunity”. In other words, wide-scope modality should not be solely associated with epistemic modality in the classification of modal meaning as it can be observed in present-day English.

An important contribution to the study of the recent development in the English auxiliary domain has been made by Krug (2000). He offers an analysis of so-called emerging modals and shows how lexical structures are changing their categorial status and are developing features typical of the core modals in a process of grammaticalization. The following are the processes discussed and analysed by Krug:

- want to > wanta > wanna
- is/am/are going to > ’s/’m/’re going to > gonna
- have got to > ’ve got to > gotta

Ziegeler (2011) offers a more general overview of grammaticalization in the field of modality. Starting from the cross-linguistic evidence presented in Bybee et al. (1994), Van der Auwera and Plungian (1998) use semantic maps (see Chapter 25) both to capture polysemy from a synchronic point of view and to represent the semantic paths along which modal meanings have evolved; the paths are similar in languages from different families.

### 4 Conclusion and future directions

The domain of study of modality is such that any survey is unavoidably limited: morphological and syntactic realizations of modality (cf. section 1) and lexical realizations of modality other than modal verbs, and their impact on the concept of modality as such, have not been discussed. The question of how modal meaning as expressed by modals is approached in formal semantics (cf. e.g. Kaufmann et al. (2005), Portner (2009: chapter 3) for an overview; see also Chapter 4), in cognitive and force dynamic
approaches (cf. e.g. Mortelmans (2012) for an overview; see also Chapter 5) or in Construction Grammar (Bybee 2010) has not been addressed.

The scope of the chapter notwithstanding, it has become clear that the semantics-pragmatics interface is relevant to the topic in more than one way: it was pointed out in the brief historical sketch that in some cases implicated meanings conventionalized and have in this way come to constitute the semantic core of specific modal verbs. It is important to distinguish between this development and the foregrounded pragmatic meanings that take the form of an indirect speech act value in specific contexts (cf. e.g. You may want to put the table over there (modal meaning: epistemic possibility; indirect speech act: mild order)). Even though these illustrate conventional(ized) uses, the foregrounded speech act value still has pragmatic status; that is, it is implicated, and it does not, as such, supersede or replace the semantic core of the modal in question (cf. Stefanowitsch (2003) for a different point of view).

Features such as strength of the modality or source of the modality obviously impact on the semantic profile of modals, but these features are not distinctive in the sense that they result in mutually exclusive meaning classes. Determining the status of these parameters is by no means a straightforward matter.

Finally, the discussion of the monosemy/polysemy question brings up the question of the extent to which context contributes to truth-conditional content, an issue at the heart of the debate between so-called minimalists (cf. e.g. Borg (2012)) and contextualists (cf. e.g. Recanati (2010)).

As well as the need to clarify theoretical issues, one of the major challenges in this field remains that of specifying fine-grained meaning distinctions and connecting them coherently with theoretical models. Much still needs to be done.

Further reading


References

Ilse Depraetere


Modality


Ilse Depraetere


Related topics

Chapter 13, Sense individuation; Chapter 19, Lexical and grammatical aspect; Chapter 20, Tense.