



*Theory and Practice in English Studies 3 (2005):
Proceedings from the Eighth Conference of British, American
and Canadian Studies. Brno: Masarykova univerzita*

The Grammatical Nature of the English Modal Auxiliaries: a Hypothesis

Sigbjørn L. Berge

Agder University College, Kristiansand, Norway

The grammatical defectiveness of the English modal auxiliaries is well known. In standard usage they have no infinitive, no present or past participle, and no ‘-(e)s’ inflection with 3rd person singular subjects. The descriptive facts are well established, but the explanatory question is a theoretical challenge. This is the topic to be addressed here on a strictly synchronic basis. It outlines a hypothesis about the nature of the modals that will bring together under the same explanatory concept the well-known facts about gaps in the inflectional paradigm. The proposal is that the English modals can be regarded as inherently non-indicative verb forms in their lexicon representation. This hypothesis seems to offer an interesting explanatory perspective on several aspects of modal auxiliary behaviour.

1.1 The formal restrictions on the English modals are well known, and there is hardly any need for additional empirical documentation of these grammatical facts, at least not in standard present-day English (see for instance Palmer 2001: 100-01). The grammatical potential of the genuine modals (as opposed to modal auxiliary equivalents like ‘have to’, ‘be going to’ etc.) is highly restricted. In particular, the genuine modals have no infinitive form – hence they cannot follow another modal, which requires an infinitive form after it, e.g.:

[1] We will ***must / have to** tell them.

Furthermore, they have no past participle and are therefore barred from combining with the auxiliary ‘have’ to form the perfect:

[2] We have ***must / had to** tell them.

Nor do they have an ‘-ing’ participle, and as a consequence a modal is unacceptable in a position where this participle is required, e.g.:

[3] We regret ***musting / having to** tell you this.

That is, the modals have no infinitive and no participles, i.e. no non-finite forms – only finite ones. In addition, their present tense forms do not accept the regular suffix ‘-(e)s’ with 3rd person subjects (unlike the modal auxiliary equivalents):

[4] Somebody ***musts / must (has to /*have to)** tell them.

These are some highly language-specific facts about present-day standard English. In comparison with closely related languages like the Scandinavian ones these facts about the English modals are quite striking, as their cognate Scandinavian forms have at least infinitives and past participles. Also, from the point of view of historical development it is a very special situation, as the language used to have non-finite forms of the modals – and this is still the case in certain varieties of English, esp. in certain areas of Scotland and southern parts of the USA – see for instance de la Cruz (1995) for details. The morphological restrictions on the English modals are the result of a grammaticalization process. In a gradual process of auxiliarization and delexicalization they have come to look more and more like grammatical markers in present-day English. The genuine modals in their present state have gone further down this path than is the case with their cognates in other Germanic languages. The English ones exhibit a higher degree of grammaticalization in the form of tighter restrictions on their morphosyntactic potential. It is a challenging task attempting to represent the special nature of the English modals – still, an attempt to do precisely this will be outlined below.

1.2 The English modals and auxiliaries generally have been the topic of considerable linguistic research on a diachronic basis. Their historical development from Old English via Middle English and early Modern English has been investigated and analysed in considerable depth in a large number of scholarly works, in more recent years by linguists like Lightfoot, Plank, Warner, and others (see for instance Denison 1993, Part V for a survey). In a strictly synchronic description of present-day English grammar, on the other hand, we are interested in regularities and generalizations within the language system in its present state. This is the approach to be developed here. The question in focus is whether it is possible to claim any kind of explanatory sense behind the grammatical restrictions on the modals in a strictly synchronic perspective – that is: are these formal restrictions mere morphological oddities without any underlying significance, or is it conceivable that they are correlated and congruent grammatical properties?

2.1 The term ‘modality’ is used here for a category of meaning (as in Huddleston and Pullum 2002). At the heart of modality there is a semantic core to do with the speaker’s evaluation of the situation described in the sentence as a non-fact. An expression of modality contributes an element of potentiality and non-factuality to the semantic interpretation of the sentence, as opposed to the categorical factual meaning of sentences with no modal auxiliary (or any other expression of modality). This semantic distinction of factuality vs. non-factuality is also crucially involved in mood selections, understood here in the narrow sense of inflectional exponence in the finite verb only. In this narrow sense of the term it refers to a choice between indicative mood on the one hand, and imperative or subjunctive mood on the other. In present-day English this is basically a choice between indicative and non-indicative forms as there is no inflectional marker to distinguish between the imperative and the present subjunctive – or in terms of grammatical features: Mood → [± Indicative]. At a higher level of description, indicative mood may be labelled ‘the mood of factuality’, and non-indicative mood ‘the mood of non-factuality’ since a subjunctive or an imperative verb form cannot be used to represent a state of affairs as a fact.

2.2 It seems that modal auxiliaries are generally regarded as indicative in mood. This point is briefly mentioned by Quirk et al. (1985, ch. 3.52 Note):

“Verb phrases introduced by modal auxiliaries are normally classified as indicative, but it is worth pointing out that not only semantically, but syntactically, they resemble imperatives and subjunctives. They lack person and number contrast and also (to some extent) tense contrast. It follows from the lack of person and number contrast that they have no overt concord with the subject.”

As pointed out by Quirk et al., there are certain formal grammatical similarities between the modals and what we refer to as non-indicative verb forms. One significant similarity is the fact that they lack person and number contrast, which is to say that the modals do not show subject-verb agreement in the form of the verb suffix ‘-(e)s’ with 3rd person singular subjects. In this respect they are like non-indicative verb forms. Granted that imperatives and subjunctives are finite, following Quirk et al. (1985, ch. 3.52) among others, we have to restrain the subject-verb agreement rule in present-day English so that it applies to finite verb forms, but only if the mood is indicative, and only if the tense is present (except for the special case of ‘to be’). This lack of agreement is seen in the imperative if a 3rd person subject is selected, e.g.:

- [5] a) Everybody **get** (*gets) out of here!
- b) Somebody **shut** (*shuts) that window!
- c) **Don’t** (*Doesn’t) anybody move!

The same lack of agreement is also seen in the case of the verb ‘to be’ with a 2nd person subject:

- [6] Just you **be** (*are) reasonable and stick to the point!

Similarly, we see non-agreement with subjunctive verb forms, as with the preterite subjunctive ‘were’:

- [7] If I/you/he/she/they **were** elected, that would be sensational.

And we may add present tense subjunctives, as in:

- [8] a) It is important that somebody **take** care of the families.
- b) We demand that the hostages **be** released without delay.

It is debatable whether the verb forms in [8]a and b are really finite verb forms – it can be argued that they should be regarded as infinitive rather than subjunctive forms, but this uncertainty of description will not destroy the generalization about mood and subject-verb agreement: S-V agreement applies if and only if the mood is indicative.

In this respect (i.e. rejection of S-V agreement) the modals behave like non-indicative verbs. In fact, rather than regarding them as indicative forms, it may be more revealing to describe them as inherently non-indicative. It is an interesting hypothesis that the modals carry the feature [-Indicative] as an inherent mood feature in their lexicon representation. In that case a modal auxiliary may be represented in the lexicon as follows:

LEXEME (: CAN, MAY, SHALL, WILL, MUST, OUGHT, NEED, DARE)

{Lexical features}

{Grammatical features:

Verb: [+Aux]

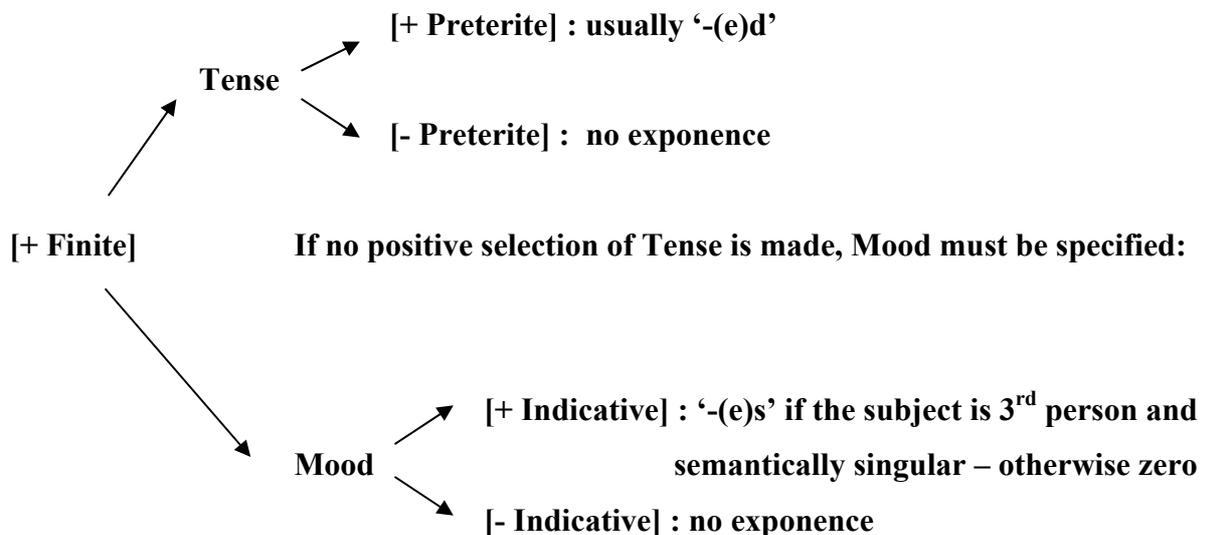
Tense:

Mood: [- Indicative]}

The morphological, syntactic and semantic evidence in favour of a hypothesis like this is worth considering. Due to limitations of space this evidence (or part of it) will have to be presented here rather cursorily.

3.1 Mood, in the narrow sense adopted here, is a category of grammar encoding modal meaning, which embodies non-factuality. This grammatical category involves a choice of inflection of the finite verb: the verb can be inflected to show indicative mood or non-indicative mood. In present-day English the scope of this particular distinction is so restricted that it is seen only with 3rd person singular subjects. Still, there is a grammatical regularity in the language to do with S-V agreement, which is bound up with selection of mood, as shown above. But if mood is a grammatical category involving finite verb inflection, it follows from the design and logic of the grammatical system that an inherent specification of mood entails finiteness. In other words, if modals are inherently specified for mood, they are by implication also inherently finite. Which is to say that their absolute rejection of the agreement marker is in fact grammatically congruent with their absolute refusal to appear as non-finite forms. The non-agreement of the genuine modals follows from the feature [- Indicative], and the higher category mood is dependent on the selection of [+ Finite] for the clause.

As regards finiteness, this is a selection which requires inflectional specification of the first verb in the verbal group for tense and mood. The special thing about English, as opposed to an inflectionally richer language like German for instance, is that a lexeme cannot carry more than one inflectional marker, as a general rule. This means that tense and mood cannot both be expressed in the same lexeme (with the exception of 'to be'). There has to be an ordering restriction on their selection, i.e.:



This analysis of finite verb inflection is developed in some more detail in Berge (to appear). In that paper the idea that S-V agreement inflection is essentially a marker of indicative mood is formalized. Also, it formalizes the restriction that the verb suffix ‘-(e)s’ is not basically a matter of present tense but presupposes and is dependent on the logically prior selection of the value [-Preterite] under Tense. Together with the representation of modal auxiliaries shown above this will serve as a condensed outline of the grammatical assumptions behind the proposed hypothesis about the nature of the modal auxiliaries.

3.2 Grammaticalization of a semantic category embodies a high degree of predictability of a certain form or structure in the sense that if one particular choice is made, another will more or less automatically also be made. For instance, if a countable noun is selected, some kind of specification of the noun will also be selected in the form of a determiner and/or number inflection. Forced, obligatory choice of form is typical of a grammaticalized semantic category. This is the case with the modals in English, but only in certain special constructions. In a conditional sentence there is a requirement for a modal auxiliary in the clause expressing the consequent if the intended meaning is non-factual or counter-factual. “In Present-day English the apodosis of a remote conditional must contain a modal auxiliary” (Huddleston and Pullum 2002: 199 – cf. ch. 14.2.2 for the same topic). This means that a sentence like [9]b without a modal auxiliary in the apodosis will be unacceptable in present-day standard usage:

- [9] a) If that were the case, the government would/might have to resign.
b) *If that were the case, the government had to resign.

The point is that in a so-called subjunctive conditional (or what we may want to call a ‘non-indicative conditional’) a genuine modal is required in the main clause in English, as opposed to the parallel situation in the Scandinavian languages, where the modal is an optional selection. Similarly, in earlier forms of English the presence of a modal in a construction of this sort was an optional matter (cf. for instance Denison 1993: 312-14). The same kind of obligatoriness is also the rule in counter-factual conditional sentences, i.e.:

- [10] a) If he had been in, I would/should/could have heard him.
b) *If he had been in, I had heard him.

Again, the modal in the clause expressing the consequent seems to be an automatic selection if the speaker intends to express a non-factual or counter-factual condition. The modal seems to operate as an obligatory grammatical marker of the meaning ‘non-fact’ or ‘contrary to fact’. In the light of the hypothesis about the modals as non-indicative verb forms this situation in present-day English is not unmotivated: these verbs are claimed to be grammaticalized markers of non-indicative mood and as such are designed for this type of expression. They correlate with the non-indicative ‘were’ in sentence [9]a above. In the case of indicative conditionals (or ‘real conditions’, ‘open conditions’) we note that the non-indicative ‘were’ is unacceptable and also that the modal is not an automatic, forced choice, i.e.:

- [11] a) If he was (*were) in, he would/could be working in his room.
b) If he was (*were) in, he was working in his room.

In these cases, where factuality is not excluded, only the indicative ‘was’ is acceptable, and the presence of a modal is not obligatory. On the assumption that the modals are inherently non-indicative verbs we can begin to make some sense of these grammatical facts. In constructing a conditional expression, the speaker has to make a decision about the factual status of the situation described in the sentence. If the state of affairs reported is considered to be most likely not the case or definitely not the case, then it is an expression of non-factuality, and markers of non-indicative mood are called for. Hence, non-indicative verbs, i.e.

subjunctive ‘were’ and a modal auxiliary, are built into the expression for the meaning ‘non-fact’ or ‘contrary to fact’.

3.3 It is assumed here that modals are specified for tense, as seen from the suggested representation of the modals shown above. But if the modals are specified for tense, we are faced with the notorious problem of accounting for the indirect tense/time relationship that is so typical of the English modals, i.e. preterite tense with non-past time reference (cf. Warner 1993: 9, for instance). Treating the modals as inherently non-indicative may serve to throw some explanatory light on this situation, however. We know from a closely related language like German that a preterite subjunctive verb form is semantically very different from the corresponding preterite indicative with respect to time reference. The preterite subjunctive is in itself not a past time expression and is therefore unacceptable with a past time adverbial in the following case:

[12] Wenn er **jetzt /morgen** /***gestern** käme, wäre es gut (from Andersson 1994: 1)

The relevant point for our purpose is that the selection of subjunctive mood has the effect of cancelling the typical semantic contributions of tense – that is, the time-referring function of tense will be void if the verb is specified for subjunctive mood. It appears that this generalization can be extended into English, on the assumption that the modals are non-indicative verb forms. Preterite tense modals like ‘would’ and ‘could’ may refer to past time, but also to present or future time, and other preterite forms like ‘might’ and ‘should’ have no past time use at all (at least in simple sentences). This may look like a rather messy area of English grammar, as it seems to contradict the very nature of tense as a matter of time reference. Still, it may not be so strange and unusual after all, if we consider the cross-linguistic evidence – the typical thing in fact seems to be a mismatch between tense form and time meaning if the mood is subjunctive. This is of course also the case with the overt subjunctive ‘were’ in English: it is preterite in form, but is not used with past time reference (according to most authorities). The interaction of tense, mood and time reference should definitely be dealt with in more detail, but there seems to be some relevant cross-linguistic evidence in this area suggesting that the English modals are very much like subjunctive verb forms in this particular respect. It will be counted in favour of the hypothesis proposed above: if the modals are inherently non-indicative, do not expect the preterite tense to be capable of expressing the same time meaning as it does with verbs in the indicative mood.

4. It is difficult, if at all possible, to provide hard and fast evidence for or against the hypothesis outlined above. Rather, it will be a matter of finding evidence that can be used to illustrate whether this idea makes better sense than alternative ones or not – whether it seems more plausible or not. This may be based on language-internal observations as well as cross-linguistic ones. As regards further language-internal evidence, there are certain well-established facts about modal auxiliary behaviour in English that may be interpreted in an interesting way granted the idea that the modals are grammaticalized as non-indicative verbs. One case in point may be the so-called marginal modals ‘dare’ and ‘need’. Generally speaking, these verbs behave as genuine modals only if the meaning of the construction they occur in is non-assertive (see e.g. Quirk et al. 3.41-42 for details). The question is whether this restriction amounts to anything more than a grammatical oddity and a meaningless irregularity. We may note that the notions of non-assertiveness and non-indicativeness are semantically related: they are notions of non-factive meaning. But if the genuine modals are in fact inherently specified for non-indicative mood, as suggested above, the restriction on ‘need’ and ‘dare’ in the modal auxiliary use is semantically congruent with their lexicon representation as non-indicative verbs, i.e. the language avoids a non-indicative verb in an

assertive context if a synonymous alternative is available. In this perspective the restriction on the marginal modals may seem to be something more than just a meaningless irregularity. However, this begs the question why this restriction does not apply to all the genuine modals, which are all claimed to be inherently non-indicative. Somehow it may be possible to reconcile these facts, but at present this issue will have to be left open for further elaboration and conceptual analysis.

Also, the relevance of cross-linguistic observations should be elaborated. As regards the presence and absence of a modal auxiliary in conditional sentences dealt with above in 3.2, this will be an interesting topic for contrastive investigations. A comparison of modal auxiliary occurrence in English and Norwegian will show that a finite modal is often an optional selection in Norwegian, but obligatory or close to obligatory in the corresponding English sentence. This difference of modal auxiliary presence is something that calls for some kind of explanation. It may well be that the hypothesis outlined above about the nature of the English modals will pave the way for a plausible analysis of certain such cross-linguistic observations that will otherwise be left unaccounted for. Although this is something that cannot be followed up here, it may be an approach to mood and modality across languages that will open up some fruitful explanatory perspectives.

References

- Andersson, S.-G. (1994) 'Proximität und Distalität im deutschen Tempus/Modussystem' *Nordlyd* 22: 1-7, Tromsø: University of Tromsø.
- Berge, S. L. (to appear) 'The grammatical identity of the English verb suffix -(e)s' *Proceedings from The Ninth Nordic Conference for English Studies*. Aarhus: University of Aarhus.
- Cruz, Juan de la (1995) 'The geography and history of double modals in English: a new proposal' *Societas Linguistica Europaea* XVI: 75-96.
- Denison, D. (1993) *English Historical Syntax: Verbal Constructions*, London: Longman.
- Huddleston, R. and Pullum, G. (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Palmer, F. R. (2001) *Mood and Modality* (2nd edition), Cambridge: Cambridge University Press.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*, London: Longman.
- Warner, A. R. (1993) *English Auxiliaries: Structure and History*, Cambridge: Cambridge University Press.