Lightfoot, D. W.

Principles of Diachronic Syntax

Chapter 2:

A paradigm case: the English modals

Cambridge: Cambridge University Press, 1979
2 A paradigm case: the English modals

In this chapter I shall seek to establish that grammars can undergo radical re-structurings in the course of time, and I shall introduce the Transparency Principle, which will account for when such re-structurings should occur. The argument will be that the grammar of NE, unlike that of OE, has an initial structure syntactic category of ‘modal’, and that the development of that category sheds much light on the nature of syntactic change.

2.1 Modals in Modern English

As a prelude to the diachronic argument, I shall discuss the status of modals in NE and claim that they should be analysed as originating in the same S as that in which they end up in surface structure. This might be taken as a controversial position since there has been a widespread belief that auxiliaries, and more particularly modals, are to be derived transformationally from higher predicates. This view originated with Ross (1969a), who presented ten arguments that auxiliaries and verbs ‘are really both members of the same lexical category, verb’, and two arguments purporting to show that they must be main verbs. Since this analysis has achieved a certain popularity, it is worth examining with some care. §2.1.1 will show that all of these arguments are badly flawed as they stand, and §2.1.2 will show that there are good reasons to derive modals from within the same S as the verb they ‘govern’.

2.1.1 To illustrate Ross’ position, windows may be broken by rioters is to be derived from one of two initial structures, (1) or (2), which happen to correspond to the two possible semantic readings, i.e. the permission (root) or possibility (epistemic) readings. In each case the modal may originates in a higher predicate to the rest of the sentence. Max was chortling would come from an initial structure like (3).
2 A paradigm case: the English modals

(1)

S
  NP I
  VP
    V
    NP
      [+V
      [+Aux
      [+Modal
      +may
    N it
    S
      NP rioters
      VP
        V
        NP windows

(2)

S
  NP
    S
      NP rioters
      VP
        V
        NP windows

(3)

S
  NP Max
  VP
    V
    NP
      [+V
      [+Aux
      +be
    at
    NP
      N it
      S
        NP Max
        VP
          V
          chortle
As the first argument for such initial structures Ross points out that
the term T (\(\{ \text{have} \})\) is mentioned in at least three rules of English:
Subject-Auxiliary Inversion, Negative Placement, and VP Deletion.
He notes that this is a very odd term, not even being a constituent,
and that the standard theory makes the claim that the items mentioned
in this term have no similarity which would predispose them to function
together. With no more argumentation he concludes: ‘I suggest that
this term should be replaced in all rules which mention it by the
equally natural constituent... \(\{\text{have} \}\).’ This is not an argument for
anything, but a suggestion for a new labelling convention; furthermore,
Ross’ feature analysis is essentially the same as the one in Chomsky
(1957), except for notation. The ‘problem’ is that the standard theory
has no single label to cover tense and the first auxiliary verb. Ross
tries to invent one, but it is not clear how this simplifies the grammar or
enables it to capture more generalizations. One price to be paid is that
the affixes -en and -ing will now need to come from quite a different
source, presumably introduced by special rules of complementation.

It is curious to note, parenthetically, that the fact that T (\(\{ \text{have} \})\) is
mentioned in several rules was taken as a positive argument in Chomsky
(1955). Under that theory it was assumed that a transformational
grammar consists of (i) a list of structural descriptions, (ii) a list of
structural changes, and (iii) rules associating them. Thus if a single
structural description appeared in many rules, this was a generalization.

There is, in fact, another solution which does not resort to new
labels, higher predicates or special complement types. The proposal
involves the phrase structure rules of (4). Some justification for this
unusual division is that have + en and be + ing can occur in infinitives
and gerunds but T and modals cannot (5).

(4) \(S \rightarrow NP \ Aux \ VP\)
    \(Aux \rightarrow T \ (M)\)
    \(VP \rightarrow (have-en) \ (be-\text{ing}) \ V...\)

(5) a. Sam believes Bill to have arrived at noon
    b. Sam believes Bill to be arriving at noon
c. *Sam believes Bill to arrived at noon  
d. *Sam believes Bill to can arrive at noon  

The proposal includes a rule of have-be Raising, which moves have or be from the VP into Aux position.  

\[
\text{have-be Raising} \\
X \ T \ \left\{ \begin{array}{l}
\text{have} \\
\text{be}
\end{array} \right\} \ Y \ \Rightarrow \ \text{oblig} \ \Rightarrow \ 1 \ 2 + 3 \ 4
\]

Notice that the rule will not apply if there is a modal intervening between T and \( \left\{ \begin{array}{l}
\text{have} \\
\text{be}
\end{array} \right\} \). This means that, after the application of have-be Raising, Aux will be able to consist only of T, T M, T have or T be. Therefore, Subject-Auxiliary Inversion and Negative Placement need mention only the node Aux, and VP Deletion will mention only VP; there will be no need for the T \( \left\{ \begin{array}{l}
\text{M} \\
\text{have} \\
\text{be}
\end{array} \right\} \) configuration.  

Jackendoft (1972: pp79ff) makes extensive use of this framework to capture a series of generalizations about the syntax and semantics of adverbs.¹  

Ross’ second argument is to the effect that the copula be should be analysed as a true verb and be assigned the feature [+Aux]. It behaves like a real verb in that it occurs in the same position in the sentence and it undergoes Gapping. On the other hand it behaves like an auxiliary in that certain quantifiers hop over it (6).  

(6) a. they \{all \} both are handsome \{each\}  
   b. they are \{both \} handsome \{each\}  

¹ Jackendoft gives as further evidence the sentences  
(i) for John \{not to have \} left disturbs me  
(ii) for John \{not to be \} the man I’m looking for disturbs me  

Assuming with Chomsky (1973) that infinitival complements are tenseless, the structural description for have-be Raising would not be met and so the negative would be placed before the have or be since it is still in the VP. I am less than happy with this argument because (a) the grammaticality judgments are not clear, and (b) by most formulations to replaced Aux, therefore the predicted order would be the less acceptable \( \ldots \) to not have left, instead of \( \ldots \) not to have left.
It is true that the main verb be and the auxiliary be share certain properties (e.g. both undergo Subject-Auxiliary Inversion), but it is curious to call the main verb be an 'auxiliary' in order to capture these similarities. Moreover, it is scarcely an argument to label the copula \[ \begin{array}{c} +V \\ +Aux \end{array} \] if one shows that it shares properties (word order and Gapping) with 'true verbs', which are \[ \begin{array}{c} +V \\ -Aux \end{array} \], i.e. properties not shared by 'verbs' which are \[ \begin{array}{c} +V \\ +Aux \end{array} \].

The third argument is based on a putative constraint that 'no agent in a for-to or poss-ing complement can be identical to the subject of a higher sentence, as long as only for-to or poss-ing sentences intervene'. This is to account for the ungrammaticality (for Ross) of (7b).

(7) a. he forced me to be examined by Dr Hito
   b. *he forced me to be examined by him

Unfortunately many people do not find (8) to be any better than Ross' (9), to which he assigns ??. (8) has a that complementizer intervening between the subject I and the agent me.

(8) I want Mary to tell Tom that Peter got Bill to try to force Jack to be examined by me
(9) I want Mary to convince Tom to get Peter to try to force Jack to be examined by me

So the constraint does not seem to be well-founded. However, Ross assumes its correctness and uses it as evidence that the may of permission is a true verb which has a first person subject (when used as a performatif) and undergoes a rule of 'Flip'. Such a constraint would then explain the difference between (10a) and (10b), since they would have essentially the same source as (11a) and (11b).

(10) a. you may gladly be examined by Dr. Hito
   b. *you may gladly be examined by me
(11) a. I gladly allow you to be examined by Dr. Hito
   b. *I gladly allow you to be examined by me

Ross concludes that 'the rule of Flip applies only to verbs – hence the fact that it must apply to may argues that this modal is also a verb'. The illogic of that needs no comment. Moreover the ungrammaticality of (10b) and (11b) has more to do with the adverb gladly than the
interaction of the putative constraint and Flip. If *gladly* is deleted the sentences become perfectly well-formed. It is somewhat unusual to use *allow* as a performative but *(I’ll allow)* you to be examined by me is quite acceptable. Therefore, the constraint must be abandoned and this argument for treating *may* as a higher verb is no longer valid.

As the fourth argument, Ross points out that *be-ing* and *have-en* can occur under verbs such as *seem*, which normally require the next lowest verb to be [+stative], and not under verbs such as *force*, which require the next lowest verb to be [−stative]. These facts can be captured by treating *be* and *have* as ‘true verbs, with the features [+Aux] and [+stative].’

(12) I forced him to

\[ \begin{array}{c}
\text{learn the answer} \\
\text{*know the answer} \\
\text{*be sleeping} \\
\text{*have slept}
\end{array} \]

(13) he seems to

\[ \begin{array}{c}
\text{*learn the answer} \\
\text{know the answer} \\
\text{be sleeping} \\
\text{have slept}
\end{array} \]

However, the data are more complicated. One can, in fact, have [−stative] verbs immediately under *seem* and [+stative] verbs under *force* in certain circumstances, for example, if *seem* is in the past tense or if the complement S contains a plural NP or an indefinite adverbial clause.

(14) a. Bill seemed to learn the answer
    b. Bill seems to learn the answers
    c. Bill seems to kick Floyd whenever he has a chance
    d. Max forced Harry to be washing up when his mother came

Chomsky (1972b) points out that there is no need for any special initial structure constraints on the complements of *seem* and *force*; it is always possible to assign some interpretation. Thus any well-formed sentence can be embedded under *seem*, but no ill-formed sentence: *John reads a book* is unacceptable except as a stage direction, etc., and likewise *John seems to read a book.* Compare *John is reading a book*, which is acceptable, and likewise *John seems to be reading a book.*

The next three arguments are all based on pronominalization, in particular on the assumption that the antecedent of a pronoun must be
a constituent. Ross assumes that so is a pro-S (as in *I hope that we will win in Vietnam, but no sane man hopes so*) and then shows that so may replace singing, been singing, or have been singing.

\[(15) \quad \text{they said Tom may have been singing, and} \\
\quad \left\{ \begin{array}{l}
\text{might have} \\
\text{so he might have been} \\
\text{might}
\end{array} \right. \]

Ross proposes that the derived structure of the embedded sentence be \[(16)\), so that in all cases there is an appropriate antecedent for so, namely a S.

\[(16)\]

In the other pronominalization arguments Ross is concerned to provide a NP antecedent for *which* and *that* in sentences like \[(17)\) and to derive \[(18)\) by the rule of S Deletion, where *it* replaces a S which is identical to some earlier S.
(17) a. they said that Tom likes ice cream, \( \{ \text{which he does} \} \) 
    \( \{ \text{and that he does} \} \) 

b. they said Tom might have been sleeping, 
    \( \{ \text{which he might (have (been))} \} \) 
    \( \{ \text{and that he might (have (been))} \} \)

(18) Max was shorting when I got up yesterday morning and he was 
    still at it when I went to bed that night

At the time that Ross’ article was written, many arguments for in- 
creasingly abstract underlying syntactic structures were being based on 
the assumption that an antecedent for Pronominalization should be a 
constituent at the time that Pronominalization applies. This assumption, 
of course, was quite a priori. It has since been shown that there are 
many cases where pro-forms refer back to something which is not a 
constituent at any point in the derivation.\(^1\) Examples to counter Ross’ 
arguments here would be (19)-(21).

(19) Tom refused the hot dog, \( \{ \text{which} \} \) wouldn’t have happened 
    with filet mignon

(20) Tom refused the hot dog, and it wouldn’t have been so with 
    filet mignon

(21) Tom refused the hot dog, and it happened again with filet mignon

Two arguments concern S Deletion and word order in German. It 
is difficult to see what bearing these arguments have on the analysis 
of auxiliaries in NE. It is entirely likely that in certain languages 
English modals will be translated by truly verbal forms, but one cannot 
argue that on universal grounds the English modals are therefore 
verbs.

There is an interesting argument that \textit{windows may be broken by rioters} should be derived from either a higher transitive verb \textit{may} as in 
(1) (which gives the permission or ‘root’ reading) or an intransitive 
as in (2) (which gives the possibility or ‘epistemic’ reading). Ross 
shows that the sentence can be disambiguated by the adverbs \textit{gladly} 
and \textit{possibly} and then points out that only the \textit{possibly} sentence can 
undergo \textit{there} Insertion. (Chomsky (personal communication) points

\(^1\) This is based on the usual assumption that in \textit{John refused the hot dog} the subject 
and verb do not form a constituent. Chomsky (1972b) cites \textit{John turned down the} 
hot dog \textit{flat, but it (that) wouldn’t have happened with filet mignon, and half the class} 
\textit{flunked physics, which would never have happened in Eng. Lit.}
out that, in fact, the distribution of gladly and possibly is not a function of the root/epistemic distinction, since the same incompatibility arises with were instead of may be.)

(22) *there may gladly be windows broken by rioters
(23) there may possibly be windows broken by rioters

(23) would be derived by applying Passive and there Insertion on the lower S and then raising there on the next cycle by Subject Raising. This kind of derivation would be needed independently for sentences like there seems to be an octopus lying on the quay, since there is good evidence that seem is an intransitive verb. However, Ross’ argument does not go through because, as he shows, (22) is also derivable by his theory, an unhappy result. (22) would be derived from an initial structure like (1) by applying Passive and there Insertion on the lower cycle, raising there on the next cycle and then moving there into subject position by Flip. The transitive may must be able to undergo Flip in order to derive Max may gladly be examined by Dr. Hito (in the permission reading); recall the discussion of Ross’ third argument (p85).

A further argument is based on the fact that (24) and (25) ‘are felt to be variants of one another’ but that in the Aspects analysis they must come from quite different initial structures, namely (24a) and (25a) (Ross’ trees).

(24) Ella doesn’t need to go

(24) a.
2 A paradigm case: the English modals

(25) Ella need not go

(25) a.

S

neg

NP

Ella

VP

Aux

T

pres

M

need

V

go

Ross, however, analyses (24) and (25) as derived from the same initial structure, thereby capturing their similarities.

(26)

S

NP

it

V

VP

[+V

+Aux

+pres

+neg]

NP

Ella

VP

[+V

-Aux

-M

need

it

S

NP

Ella

VP

V

go

An extra transformational rule is then needed to change the features under certain circumstances. Ross formulates it roughly as (27).

(27) dare opt need \[\Rightarrow [+\text{Aux}][+\text{Modal}] \text{/negative contexts}\]

If this rule fails to apply, need is treated as a real verb, and will undergo Number Agreement, do Support, normal complementation rules, etc. But if (27) does apply, then need will be treated as a modal, i.e. will not trigger do Support, will undergo Inversion in questions, allow a negative to be placed on its right, etc. The new [+Aux] need will be marked to undergo Equi-NP-Deletion obligatorily. This will necessi-
tate a new kind of marking convention, since this feature complex is introduced transformationally and therefore we cannot take the familiar course of using rule features in the lexicon. Furthermore, the [+Aux] need will also have to be marked to undergo a special rule of to Deletion to avoid *need he to be so cruel? Thus Ross will need extra mechanisms to account for the different syntactic behaviour of the two needs. It is scarcely clear that this constitutes an improvement in the grammar, but given that there are two needs (in Ross’ framework one being [+Aux] and the other [−Aux]) and that they have different syntactic properties, it is not unreasonable for the standard theory to posit the two initial structures. The standard theory was concerned with syntactic behaviour as a criterion for initial structures. If one wants to argue that (24) and (25) are synonymous and therefore should come from the same source, one has changed the criteria for positing underlying structures. That was not Ross’ position in this paper.

Ross’ final argument is based on Greenberg’s (1966) observation that in languages whose predominant order is SOV, any auxiliary will follow the verb, whereas in SVO languages an auxiliary will precede the verb. Ross claims that if one treats auxiliaries as main verbs, in Bill was writing a letter ‘the fact that was precedes its object, writing a letter, is the same as the fact that writing precedes its object, a letter, and these facts need only be stated once’.

(28)

 Presumably Ross is re-interpreting Greenberg’s universal as a fact about initial structures or ‘basic’ word order. In which case, it will be a problem for him that several of his ‘auxiliaries’ are intransitive verbs
in initial structure, e.g. *may* in (2), and only after the operation of Subject Raising and Extraposition do they end up to the left of their ‘object’. If on the other hand he interprets the universal as a fact about surface structure, then it is a claim about ‘predominant’ word order and has only the status of a statistical tendency, and it is hard to interpret his claim that treating auxiliaries as main verbs ‘explains’ the universal.

2.1.2 Having been very negative and shown the faults in the internal logic of Ross’ arguments for deriving modals from higher predicates, we turn now to an interesting set of generalizations which can be captured by deriving modals from the same *S* as they end up in in surface structure. I shall first outline some of Jackendoff’s proposals concerning ‘speaker-oriented adverbs’. Jackendoff (1972: ch.3) distinguished four classes of *-ly* adverbs: speaker-oriented, subject-oriented, manner, and finally adverbs like *merely*, which do not concern us here. Speaker- and subject-oriented adverbs occur only in *S*-initial and Aux position or *S*-final with pause. Speaker-oriented adverbs often have a paraphrase *S* is *Adj* to *me*, *I am* *Adj* that *S* or *I consider S Adj* (e.g. *happily*, *John sang the aria*; cf. the subject-oriented *carefully John drove his car through the crowded intersection*. Each of these is distinct from the manner interpretation in *John sung the aria happily* and *John drove his car through the crowded intersection carefully*). Jackendoff proposed that a rule of semantic interpretation, *P* _speaker_, assign a semantic structure ADJ (SPEAKER, f(NP¹, ... NPⁿ)). If *S* is a sentence containing a non-strictly sub-categorized adverb, f(NP¹, ... NPⁿ) will represent ‘its functional structure, that is, the relation between the verb and the strictly subcategorized arguments NP¹, ... NPⁿ’. ADJ is the semantic content of the adjectival counterpart of the adverb undergoing the rule. The rules for subject-oriented and manner adverbs are stated by Jackendoff as follows:

**P** _subject:_ If Adv₁ is a daughter of *S*, embed the reading of *S*... as one argument to Adv₁, and embed the derived subject of *S* as the second argument to Adv₁.

**P** _manner:_ If {Adv} is dominated by VP, attach its semantic markers to the reading of the verb without changing the functional structure.

These rules give minimal semantic structures as (29) and (30) respectively.
(29) \( \text{ADJ} (\text{NP}^i, f(\text{NP}^i, \ldots \text{NP}^n)) \) (where \( 1 \leq i \leq n \))

(30) \( \left[ \frac{f}{\text{ADV}} \right] (\text{NP}^i, \ldots \text{NP}^n) \)

'...adverbs will be marked in the lexicon as to which of the possible semantic structures they can enter into. For example, *certainly will be a predicate over an S; the reading of *happily in happily, Frank is avoiding us will be a predicate over an S and the argument SPEAKER'. Then, if *evidently, a speaker-oriented adverb, is in final position without pause (i.e. dominated by VP), as in *John walked in evidently, only \( P_{\text{manner}} \) will be applicable, giving \( \left[ \frac{\text{WALK}}{\text{EVIDENT}(S)} \right] (\text{JOHN, IN}) \), which is ill-formed because of the unfilled variable S. With *evidently in initial position, \( P_{\text{speaker}} \) applies giving the correct interpretation: EVIDENT to SPEAKER (WALK (JOHN, IN)). \( P_{\text{speaker}} \) operates on adverbs dominated by S. The predicted positions in surface structure for adverbs dominated by S are all the possible major constituent breaks. Assuming the base rules of (4), particularly that analysis of Aux, these positions are: initial, before Aux, after the first auxiliary and final with pause.

(31) a. *evidently John could have left before noon
    b. John evidently could have left before noon
    c. John could evidently have left before noon
    d. *John could have evidently left before noon
    e. *John could have left evidently before noon
    f. John could have left before noon, evidently

There are two phenomena which will take on importance for the analysis of modals. Firstly, Jackendoff points out the apparent incompatibility of S adverbs and Subject-Auxiliary Inversion.

(32) a. *did Frank probably beat all his opponents?
    b. *what has Charley evidently discovered?

VP adverbs are fine in these contexts and so are the higher predicate paraphrases of these S adverbs and tag questions.

(33) a. did Frank easily beat his opponents?
    b. what has Charley suddenly discovered?

(34) a. is it probable that Frank beat all of his opponents?
    b. what is it evident that Charley has discovered?

(35) a. Frank probably beat all of his opponents, didn’t he?
    b. Charley evidently discovered a new way, didn’t he?
This incompatibility holds not only with Inversion marking interrogatives but also with Inversion conditioned by optionally preposed elements.

(36) a. Bill (apparently) has never seen anything to compare with that
b. Tom (probably) ran so fast that he got to Texas in ten minutes

(37) a. never has Bill (*apparently) seen anything to compare with that
b. so fast did Tom (*probably) run that he got to Texas in ten minutes

One way to handle this would be to put identical restrictions on *wh Movement, Negative Adverb Preposing and so Adverb Preposing, prohibiting them just if there is a S adverb present. This is very uneconomical and misses the point that the generalization seems to involve Inversion. We cannot require the absence of any S adverb in the structural description of Subject-Auxiliary Inversion, because suppressing the inversion makes things even worse.

(38) a. *never Bill apparently has seen anything to compare with that
b. *so fast Tom probably ran that he got to Texas in ten minutes

Jackendoff argued that Inversion introduces a semantic factor not present in the non-inverted forms, which is incompatible with the reading of S adverbs. (The problem may be more general than this, since *who evidently discovered uranium seems to fall into the same class, although not involving Inversion.)

Secondly, there cannot be more than one subject-oriented adverb, and that one must be the last S adverb in the sentence (excluding final position with pause).¹

¹ Jackendoff treats adverbs of frequency, e.g. often, never, as S adverbs (see (39)). But Lydia White has pointed out that such adverbs are not incompatible with Inversion (e.g. (37a), and never (before) has he been late). This may stem from a failure to isolate this set of adverbs from other S adverbs, or it may indicate that S adverbs are not, in fact, incompatible with Inversion. Notice that (38a), which is alleged to indicate that suppressing the inversion with S adverbs makes the sentences even worse, involves one S adverb which is an adverb of frequency (never) and one which is not (apparently); if apparently is omitted, the sentence is still ungrammatical, and therefore is irrelevant to the claim that S adverbs and Inversion are incompatible.
Now, a minimal requirement for any treatment of English modals is that it be able to capture the difference between the 'root' and 'epistemic' senses of each modal. Ross' treatment allowed such a distinction on the basis that root modals came from higher transitive predicates and epistemics from higher intransitives (see (1) and (2)); thus he drew the root/epistemic distinction at the level of initial syntactic structure. This involves various problems, as I have shown, and ignores the almost identical syntax of the two types of modal.\(^1\) The

\(^1\) The root/epistemic distinction is probably cross-linguistic and independent of variations in syntax. I know of two arguments that the two classes of modal have a different syntax, neither very strong. Ross (class lectures, 1968) claimed that the negative could contract only with root modals.

**John mustn't do the shopping tomorrow**

*John mustn't have been a student*

Not many people agree with Ross on these judgments and, in any case, this rule can't/couldn't be simplified any further is well-formed despite the epistemic modals.

Another argument is that the deleted must in **John must do the shopping and be a student** can only be read in the same way as the first must; the sentence cannot be read as 'John is required to do the shopping and he is certainly a student'. Thus the syntactic deletion rule must 'know' whether both modals are epistemic or root; if one is epistemic and the other root the rule will not apply. This argument is unsound because in the undeleted **John must do the shopping and must be a student** again the modals must be read in the same way. Kuno (1972) has convincingly shown that this kind of thing is not dictated by the grammar but is due to a general perceptual strategy which assigns parallel interpretations to adjacent ambiguous objects. Each of the objects A and B is ambiguous in terms of which is the nearest surface, but when juxtaposed they will generally be interpreted in the same way.
claim made here is that if one derives modals from the same S as they end up in, specifically from an Aux node immediately dominated by S and dominating only T (M), the epistemic readings can be derived by an independently motivated rule, namely Jackendoff’s rule for speaker-oriented adverbs. Auxiliaries and S adverbs are in the same syntactic configurations, dominated directly by S and the semantic structures of epistemic modals are identical to those for speaker-oriented adverbs, as advocated by Jackendoff. The scope of epistemics, unlike roots, is the whole of the S in which they occur1 and epistemics all indicate some degree of possibility. Therefore a minimal semantic structure for epistemic modals, as in *John \{must \} be here*, will have to include something like POSS(JOHN BE HERE). Features indicating the degree of possibility will then be added depending on the particular modal. This semantic structure is identical to that postulated by Jackendoff for speaker-oriented adverbs, and, since the syntax is the same, Jackendoff’s P\_\text{speaker} can easily be generalized to epistemic modals. The really convincing evidence for this position comes from the fact that exactly the same conditions obtain for the rule interpreting epistemics as for the one interpreting speaker-oriented adverbs. This suggests that there is probably just one rule involved.

Firstly, if a modal has undergone Subject-Auxiliary Inversion, it cannot be interpreted epistemically. That is, epistemic modals and Subject-Auxiliary Inversion are incompatible, although there is some dialect variation. For many speakers (40a) and (40b) are ambiguous in that the modal may have root or epistemic sense, but in (41) the modals can have only the root reading:

(40) a. Max \{may \} work at the McGill library
    b. Max \{may \} drive at 90mph

(41) a. \{may \} Max work at the McGill library?
    b. \{may \} Max drive at 90mph?

1 Interestingly, with the exception of speaker-oriented adverbs, which are outside the scope of the modal. *Evidently John must be a student* means ‘it is evident that it must be the case that John is a student’ and not ‘it must be the case that it is evident that John is a student’.
Presence of *have + en usually forces an epistemic reading, so if our explanation is right, it should not be possible to invert a modal if there is a *have + en in the same S.

(42) a. \{Max must
\{must Max\}

\} have studied at McGill

b. \{Max must
\{must Max\}

\} have driven at 90mph?

(43) a. Max may never have seen such a show

b. *never may Max have seen such a show

This incompatibility is of the same status as the incompatibility of Inversion and speaker-oriented adverbs. Parallel to the adverbs, tag questions allow epistemic readings and Inversion can be done on the non-modal paraphrases.

(44) a. Max must drive at 90mph, mustn’t he?

b. is it possible that Max studies at McGill?

Secondly, like speaker-oriented adverbs epistemic modals can precede but not follow subject-oriented adverbs (46), although there are no such ordering restrictions with speaker-oriented adverbs (45).

(45) a. \{John will evidently
\{evidently John will\}

\} have opened the secret door by now

b. \{Bill should probably
\{probably Bill should\}

\} have left by now

John will slowly

(46) a. ?? slowly John will \{have opened the secret door by now
?? John slowly will

Bill may quietly

b. ?? quietly Bill may \{have left by now
?? Bill quietly may

What I hope to have shown in this section is that in a transformational grammar of NE, there is good reason to derive modals from the same S as the one they end up in and not to treat them as underlying verbs. The arguments in the literature for deriving them from higher predicates are faulty and there are significant generalizations to be captured by generating them within their own surface S. Also, the wide range of transformations needed to derive the appropriate surface structures require an extremely permissive, hence undesirable theory of grammar. Given the view of grammar adopted here, my conclusion should not be
very surprising, because modals simply do not have the syntactic properties of verbs: they do not undergo Number Agreement or do Support but do undergo Subject-Auxiliary Inversion and Negative Placement. They cannot appear in infinitives and gerunds, cannot occur adjacent to each other and cannot take normal complementation forms. No doubt there is much more that could be said about the proper analysis of modals in NE, and there have been more recent arguments supporting a Ross-type analysis (Huddleston 1974, McCawley 1975, Newmeyer 1975, Palacas 1971). Having shown the plausibility of an analysis such as (4), I turn now to the diachronic data and show that this analysis permits an explanation for some historical changes, particularly for their timing. When we have developed a theory of change, this will constitute a further reason to adopt this analysis.

2.2 Historical re-analysis

2.2.1 When we look at Old and Middle English, we find a very different state of affairs. Here verbs such as sculan, willan, magan, cunnan, motan, which we usually translate into NE with modals, have all of the characteristic properties of verbs. They have full person-number paradigms (undergoing Number Agreement) and behave like verbs with respect to Negative Placement and Inversion. (Examples will be provided when the historical changes are discussed in §2.2.2 and §2.2.3.) They can occur adjacent to each other in series, and in infinitives and gerunds. They occur in sentence-final position, as one would expect if they were verbs, since at this stage English had an underlying SOV order (see Canale 1976). They can occur with normal complementation types and many can take surface direct objects, although they cannot then be translated in NE with a modal.

(47) a. hwæt þær foregange, oððe hwæt þær effýlge, wæ ne cunnun ‘what came before, or what comes after, we do not know’. Bede

1 Most of these arguments are based in some way on assumptions that there is a one-to-one relation between meaning and deep structure, and are beside the point for the view of grammar adopted here, which permits rules of semantic interpretation and does not require a one-to-one relation between initial phrase markers and semantic representations. Such a theory does not disregard semantic facts, but it simply denies a certain theory about them.
b. she koude muche of wandrynge by the weye ‘she knew much of wandering by the way’. Chaucer

c. seþe sculde him undred denera ‘he who owed him a hundred denarii’. c.975 Rushworth Gospels, Matt. xviii 28

d. ac him sealde ða mihte se ðe mæg ealle ðing ‘but he who can do all things gave him the power’. c.1000 Aelfric, Feria vi in Prima Ebdomada Quadragesimae 157

e. ich wille, þat þou suere On auter ... ‘I want you to swear ...’. 1300 Havelok 388

f. heora non swaðeah nolde befrinian hwæt heo ðær wolde ‘none of them, however, would ask what she wanted there’. c.1000 Aelfric, Feria vi in Tertia Ebdomada Quadragesimae 62

What all this means is that for OE there would be no justification for setting up a syntactic category ‘modal’. What we translate with modals in NE, all behave exactly like ordinary, complement-taking verbs in OE. Therefore, it is entirely reasonable and apparently uncontroversial¹ to argue that the initial structure of *ne con ic noht singan* ‘I cannot sing’ Bede, Cædmon, would be (48), at least insofar as can is treated as a main verb. (I omit details such as the constituency membership of Tense, returning to that in §2.4.)

(48)

```
  S
 / \   / \   / \   / \\
Prec S NP I VP   \
   \    \  \   \  \   \  \  \\
    neg V   S    \\
           /     \\
          /      \\
         V       \\
          can    \\
          /      \\
         S       \\
          /     \\
         NP I   \\
          /     \\
          V     \\
          sing  
```

¹ Among generativists, Wagner (1969) and Allen (1975), for example, take this position. Although Traugott (1972a) analyses modals as auxiliaries, she says (p69) ‘the predictive use of *will*- and *shall*- is rare in OE, however. Both verbs were originally main verbs, and it is as main verbs that they are most commonly found’, and (p109) ‘the modal and perfect auxiliaries were originally main verbs that introduced subordinate clauses’. Visser (1963–73: §1565) says of OE *willan*, for example, that it was a ‘verb that gradually became toned down to the status of an auxiliary’; of *shall, may, can, will* and *ought* he says (§5.48) ‘originally ... they were not function words, but full or independent notional verbs that syntactically did not differ in any way from the other full verbs’. 
If our analyses of OE and NE are correct, a radical re-structuring took place in the development of NE, and when we look at how that came about there are some conclusions to be drawn about the nature of syntactic change and about the abstractness of synchronic syntactic analyses.

We have here a case of a radical change in initial structure. Diachronic syntacticians have often tried to argue that there can be no such thing as a radical re-structuring. Robin Lakoff (1968: ch.6), for example, sought to show that much syntactic change was a function of changes in lexical redundancy rules and not even of changes in transformations. Hence syntactic change took place only or mostly through changes in governed rules. Her claim was that diachronic syntax is at least easier to do if one assumes that the phrase structure rules do not change.

Secondly, Traugott (1972b) has argued that ‘there is no such thing as “pure syntactic change”..., where reference is made exclusively to syntactic rules, and not at all to either semantic or phonological factors’. Stockwell (1976) takes the same position. It seems to me that the initial structure of English modals presents us with a case of pure syntactic change, a change affecting only the syntactic component. One of the striking aspects of changes involving words such as may, must, can, should, is that they underwent very many changes in their syntax and in their meaning but that these changes seem to have proceeded quite independently of each other. This would be surprising to proponents of Generative Semantics, or to those who believe that there is a tight correlation between syntactic and semantic properties. For example, cunning (> NE can) used to mean ‘to have the mental or intellectual capability to, to know how to’, and was sometimes contrasted in the same sentence with magan (> NE may), which meant ‘to have the physical capability to’. This semantic distinction was lost and may developed a permission reading from ME onwards. At one point in its history may denoted ability or capacity not depending on outward circumstances and used to serve in contexts where we now use can, e.g. he can speak Klamath. Such semantic changes are typical of these words and seem to have taken place without affecting their syntax. On the other hand, a flood of syntactic changes came in the sixteenth century and had no noticeable affect on the meanings of the modals. We shall return to the question of ‘pure syntactic change’ in §3.2 and §5.1.
2.2.2 There are two stages to the story of the change: a number of apparently isolated changes took place early in the history of English. The net effect of these changes was to lead to a second stage, a re-analysis of cunnan, magan, etc., as a new category, ‘modal’. Postulating that this re-analysis took place in the sixteenth century accounts for a large number of surface syntactic changes which took place at the time.

The early changes which set the scene for the re-structuring were the following, (i)–(v).

(i) The antecedents of the modern modals (henceforth ‘pre-modals’) lost the ability to take direct objects. This seems to have been complete in ME, with the exception of can which was a good deal more resistant. Mun + direct object, meaning ‘to think of, remember’, was lost early. The last cases attested by the OED for the other object-taking pre-modals are those of (49).

(49) shall: the leeste ferthynge þaty men shal. c.1425 Hoccleve, Min. Poems xxiii 695

(can: yet can I Musick too; but such as is beyond all Voice or Touch. 1649 Lovelace, Poems (1659) 120

may: for all the power thai mocht. 1470 Henry, Wallace iii 396

(ii) Most pre-modals (sculan, magan, mōtan, agan and durran) belonged to an inflectional class generally known as ‘preterite-presents’. These were strong verbs where the preterite forms had taken on present meaning in pre-Germanic and for which new weak preterites had been made. This is analogous to Greek οἶδα ‘I know’ and Latin odi ‘I hate’, which also are past (more precisely, perfective) in form but present in meaning. Other members of this class were witlan ‘to know’, dugan ‘to be of value’, unnan ‘to grant’, purfan ‘to need’, munan ‘to think, remember’, benugan ‘to suffice’. The notable thing about this class is that the third person singular (being historically a strong preterite)

1 The OED gives one later reference, which it treats as an archaism of doubtful standing: by the feith I shall to god, 1530 Court of Love 131.
2 This is the last attested direct object but the OED has later references for can + PP and the idiom can skill of: that cunning Kaiser was a scholar voice, and coulde of grammarye, 1875 Kingsley, Poems, ‘Little Baltung’ 82; no skill of Musick can I, simple swain, 1710 Philips, Pastorals iv 23. Roger Lass points out that these may be conscious archaisms.
3 This is arguably not a genuine direct object but a ‘cognate accusative’. But may here clearly acts as an independent verb, so I classify it with the true direct objects.
4 I classify durran ( > NE dare) as a pre-modal. Dare behaves sometimes as a modal (dare he do it? he dare not do it) and sometimes as a normal verb (did he dare to do it? he didn’t dare to do it).
was _sceal_, _man_, _ann_, _dearr_, etc., i.e. it did not have the usual _-eð_ ending,
the antecedent of the modern _-s_. Thus the modals are conservative in
failing to undergo modern Number Agreement; there never was a _-s_
ending for these verbs (although they did have second person singular
and plural endings: _þū can̄ nost_, _wē cunn̄ non_, _wē cud̄ on_). However, the
truly remarkable thing is that all the _non-pre-_modals of this class were
lost. _Benegan_ dropped out of the language very early and the last cases
of _unnan_ and _purfan_ attested by the OED are as in (50).

(50) _unnan_: Meriadok was a man þat tristrem trowed ay; Miche gode
he him an. c.1320 Sir Tristrem 1928.

_purfan_: of þis cors ne thar not a-baffle. c.1485 Digby Mysterīs
III 1437.¹

_Witan_ survived into the nineteenth century, but not as a member of
the same inflectional class as the pre-modals. The OED says ‘the
original conjugation [of _witan_] ... presented many apparent anomalies,
and various attempts at normalization were made by means of analogi-
cal formations and irregular extension of the use of certain forms,
with the result that new infinitive and present-stem forms came into
existence ...’. _Witan_ was removed from the preterite-present class, or
at least abandoned its fricativeless present tense third person singular.
The date of this change is not clear but it seems to have happened in
ME. The OED says that _wat, what, wah̄t_ ... served as the first and
third person singular forms but I have found no examples in the OED
or elsewhere of such unlevelled third person forms after ME.

_Munan_ also survives into recent times, particularly in Scots. However,
it survives only as a modal: _an‘ if I mun doy I mun doy_, 1864 Tennyson,
The Northern Farmer (Old style, Lincolnshire dialect) xvii. It lost
its sense of ‘think, remember’ very early.

The history of _dugan_ is somewhat more complicated. Stratmann
gives its last occurrence as c.1360, but the OED and Visser show it
surviving as late as the nineteenth century, although by that time it is
very rare and confined to northern dialects. The OED points out that
the original inflection _deag_ as the singular present was supplanted in
the fourteenth century by _dow_ for the plural, the third singular be-
coming _dows_ in most cases. Similarly a levelled past tense form _dowed_

¹ The OED has one later reference from 1825. This is a dialect citation in Jamieson’s
Scottish Dictionary, ‘_ye thursten_’, _ye needed not_.

comes in in the fourteenth century. This would indicate that, like witan, dugan left the preterite-present class and conformed to the most common inflectational pattern. Therefore our hypothesis predicts that it would survive only as a main verb. Unfortunately this is an oversimplification. The history of dugan is just like that of dare and need in that it survived schizophrenically as an independent verb (e.g. I never dowed to bide a hard turn o’ work in my life, 1816 Scott, Antiquary xxiii; March grows Never dows, 1855 Robinson, Whitby Gloss), but also, albeit much less commonly, as a modal (e.g. ye may not, ye cannot, ye dows not want Christ, 1637 Rutherford, Letters (1862) vol. 1. p203; she doughtna let her lover mourn, 1724 Allan Ramsay, Tea-Table Miscellany (1733) vol. 1 p2).

One can only assume that it was an accident that in this inflectional class only the pre-modals survived. It does indeed seem remarkable that almost all the pre-modals had past forms with present meanings and that modern ‘past tense’ modals (should, would, could and, historically, must) generally have present sense. Also in other languages it is often the case that what we have analysed as auxiliaries have some kind of past tense form even in present sense. I have no idea why this should be, but it does not seem possible to define a class of modals (and therefore of preterite-presents) on semantic grounds, and furthermore we have seen that preterite-presents in OE and other languages encompass a very wide semantic range: ‘hate’, ‘know’, ‘grant’, ‘be able’, ‘think’, ‘need’, etc. We need not concern ourselves here with the reason; the crucial effect of the loss of the non-modal preterite-presents was that the pre-modals (including the now uncommon mun) became an identifiable class of verbs, with the unique characteristic that they did not have a fricative suffix for the third person singular. This property was shared by OE willan which is usually classed as an anomalous verb and not as a preterite-present. It was conjugated wille, wilt, wil(l)e, willap, willap, willap and, like the pre-modals descended from preterite-presents, had no third person singular fricative.¹

(iii) A third change concerns the preterites of the pre-modals. The preterite-presents shared with the weak (and therefore productive) verbs the feature that there was no phonological distinction between

¹ OE had the three verbs, willan, willian and wilmian, which were distinct in most inflectional parts. Willan is anomalous and the other two Class II weak verbs. Wilmian always had the n before the ending, and willian had, for example, /u willast, we willap and a preterite ic willole (cf. ic wolde for willan). Some of these forms coalesced in ME with the decay of inflectonal endings.
the preterite indicative and subjunctive except in the second person singular. Hence modern should corresponds phonologically to an old preterite indicative or subjunctive. However, the preterites seem to have become unstable from early times, perhaps as a result of competition from the subjunctive. NE must and ought, historically past tenses, never carry past sense, and the relations of shall/should, will/would and can/could are rarely based on a distinction of tense; for the most part should, would and could carry present-future sense and exist independently of shall, will and can. However, in OE the pre-modals used to have appropriate time reference corresponding to their tense (whereas in NE past time reference is expressed by he must have gone, which seems not to have occurred in OE). The breakdown in the productivity of the present-preterite relationship appears to have started quite early and the preterite and present tense forms developed uses independently of each other and the tense relationship between them was steadily eroded, possibly as a result of the demise of the subjunctive mood. This can be illustrated by some uses of might and should. For example, constructions like they may intend to do mischief, where may is epistemic and means ‘it is possible’, are common from early times, but might did not occur as a past tense variant of this until the sixteenth century and now it has dropped out again.\(^1\) An example with might is these two respectable writers might not intend the mischief they were doing, 1762 Bp. Richard Hurd, Letters on Chivalry and Romance 85; nowadays this would have to be might not have intended. Constructions such as you may be right are common from ME onwards (be as be may, I make of it no cure, 1385 Chaucer, LGW 1145) but you might be right is not distinct from it in terms of time reference. The past tense of the may of permission has become obsolete and he might do it cannot mean ‘he was permitted to do it’. The opacity of the may/might tense relationship was increased by new uses of might being introduced into the language with no corresponding use of may. An example would be the use in you might (have) shut the door properly, which is a suggestion amounting to a request or reproach. The first

---

\(^1\) This statement follows the standard handbooks. However, Peter Collins has pointed out to me the following examples, which seem to be earlier epistemic uses of might: and hida ealle sceton, suwa suwa mihte boon fij dusend wera ‘and they all sat down; it is possible that there were 5000 men’, Aelfric, Homilies of the A-S Church, Thorpe (1844–6); sume beladunge mihte se rice habban his uncyste, gif se reofiga wælde ne læge æforan his gesisse . . . ‘it was possible that the rich man had some excuse for his stinginess, if the leper did not lie before his sight’, ibid.
convincing example of this is as for these gentlemen... I think they might show a little more respect for their benefactors, 1748 Smollett, Rod. Random (Tauchn.) xlv p287.

Turning now to should: shall used to mean ‘to be under an obligation, be bound to’ and this meaning survives into modern times, although nowadays it sounds a little archaic: he shall hide himself in a bean-hole, if he remains on Scottish ground without my finding him (= ‘will have to’), 1818 Scott, Heart Midlothian xviii. However, the past tense of this with should, became obsolete in the fifteenth century: Arthour, as he scholde done, Sende lucyces body to Rome (= ‘he had to do’), c.1400 Arthur (EETS) 481. An example of a new use of should with no corresponding use of shall is that in I should like to have someone for a help, where I should like, suppose, hope, are just diffident ways of saying I want, suppose, hope. This usage seems to have come in in the seventeenth century; the first example I find is I should be glad to see you at my house, 1675 Wycherley, Country Wife (Mermaid) i i. Another example is the use of should under adjectives expressing some degree of possibility: is it possible that this should be the true Lord Foppington? 1695 Vanbrugh, Relapse (Mermaid) iv vi. This seems to have been introduced in the fifteenth century; Visser’s first example is it is impossible... that a soule that is bleendid in custom of synne schuld see be foule spot in his conceyence, c.1400 Cloud of Unknowing (EETS) lxxii 10. Even in oratio obliqua should does not always appear where required by normal sequence-of-tenses rules: we shall do it cannot be reported as we said we should do it.

What all this means is that the shall/should, may/might, etc., distinctions are not based simply on tense or mood. For NE there is good evidence that should and the epistemic could are separate modals, distinct from shall and can and not related to them on the basis of tense. Even in ME there was a certain amount of evidence beginning to accumulate for this analysis, since the shall/should distinction was not as transparent as, say, that between open and opened.

(iv) The standard handbooks usually identify two normal word order patterns for OE: SMVO in main clauses and SOVM in subordinates, where M indicates the pre-modals cunnan, sculan, etc. However, if the pre-modals were indistinct from verbs and assuming the underlying order to be SOV, one would expect further possibilities. OE had a rule which moved the verb to between the subject and object; this rule was obligatory in main clauses and optional elsewhere, accounting
for the usual SVO order in main clauses, and SOV or SVO in sub-
ordinates. In that case one would expect the following orders:

(51) main clauses: a. NP_s M V NP_o: John can drink the ale
    b. NP_s M NP_o V: John can the ale drink

subordinates: c. NP_s NP_o V M
    d. NP_s M NP_o V
    e. NP_s M V NP_o

We may assume two possible initial structure configurations, (52a)
for root modals and (52b) for epistemics. In (52a) the embedded subject
will never be realized in surface structure and in (52b) the embedded
subject would normally be raised to become the surface subject of the
highest verb, the pre-modal (not affecting the order of elements).

(52) a. NP_s [(NP_s) NP_o V]_o M
    b. [NP_s NP_o V]_s M

Therefore SOVM will emerge if V-Fronting does not apply on either
cycle (and hence will occur only in subordinate clauses); SMOV
will result from the application of V-Fronting on only the higher cycle
in (52a) or by Subject Raising and Extraposition (a special instance of
the OE scrambling rules) on (52b); V-Fronting on both cycles will
give SMVO from (52a). Each of these orders occurs frequently in
OE. I cite one example of each order from Aelfric (c.1000) (examples
drawn from Allen 1975).

(53) a. SMVO: ælc man mæg ðær geseon oðres manes gedoht ‘each
    man might there see other men’s thoughts’. Sermo
    ad Populum in Octavius Pentecosten Dicendus 558

    b. SMOV: ic ne mæg ðam almihtigan ahwar ætherstan ‘I cannot
    break away anywhere from the Almighty’. The
    Maccabees 29

    c. SOVM: se ðe hit rædan wyle ‘he who will read it’. Domenica
    Pentecosten 107

So correct predictions are made for OE by adopting an underlying
SOV order, a transformational rule of V-Fronting (obligatory in main
clauses, optional elsewhere), and, crucially, treating the pre-modals
as underlying verbs.1 However, in early ME there was a base re-

1 There is, in fact, one further possibility which I have not found attested: SVOM
in subordinate clauses. This would result from applying V-Fronting on the lower
analysis whereby English became underlyingly SVO (see §7.2 and Canale 1976) and the V-Fronting rule became redundant. So the relevant base structures for root and epistemics would now be those of (54), assuming that the pre-modals were still verbs.

(54) a. NP₈ M [(NP₈) V NP₆]
   b. [NP₈ V NP₀] M

Thus one would now expect to find only the orders SMVO for root pre-modals and SVOM for epistemics. Following the SOV-to-SVO base change, root and epistemics now occurred in different configurations, which was not the case before the change of underlying order. However, SVOM is attested only rarely and SMVO became the standard order even for epistemics. It is as if the grammar was reluctant to follow through on one of the implications of the earlier SOV-to-SVO change and did not want to distinguish syntactically the root and epistemic pre-modals, perhaps because to do so would endanger mutual communicability between the conservative and innovating generations. Any rate, if we assume, as we must, that the pre-modals were still underlying verbs, then to account for the scarcity of SVOM surface structures we must postulate a further change: a rule is needed which would have the effect of mapping SVOM structures into SMVO. A rule of Extrapolation would have the desired effect. However, Extrapolation at this time probably did not exist, as we shall show in §4.3, because infinitival subject complements were base-generated in rightmost position (and likewise finite that complements although the data is less clear on this count). Therefore there are two possible analyses: (a) to invoke a probably not otherwise needed rule of Extrapolation operating only on epistemic modals and in any case operating obligatorily, or (b) to claim that as the SOV-to-SVO word order change took place, there was another simultaneous change affecting epistemic modals such that they became underlying two-place predicates, subcategorized to co-occur with an empty subject: np may [NP₈ V NP₀].

In the latter case, in order for the empty np to be filled, NP Preposing cycle but not on the higher one, a possibility only in embedded clauses. Structures with two levels of embedding, as would be required, are not common and the absence of SVOM, if absence it is, may be accidental. However, one unexpected pattern is attested, SVMO. Fortunately this is not common and seems to occur under restricted conditions, namely when there is also a prepositional or adverbial phrase in the embedded S. I am at a loss to explain this: dat we habban moton da heofonlican wumange mid him sylfum ofre 'that we might always have the heavenly dwelling place with himself', Aelfric, Domenica v 6.
would have to apply moving the lower subject up into the subject position of the epistemic verb; the np could not be filled by a dummy it, as with similar verbs such as seem, be likely. The point is that under either analysis, following the SOV-to-SVO base change some special mechanism would be needed to distinguish the epistemic pre-modals from other one-place predicates, avoiding the expected SVOM or it M [NP . . . ]s structures: either a rule of Extraposition or NP Preposing applying obligatorily if the main verb is an epistemic pre-modal. (This discussion is predicated on the common assumption that root and epistemic pre-modals occurred in different initial structure configurations. It is not clear that this is necessarily correct. There might always have been just one (two-place) structure, with the pre-modal allowing either interpretation. If this is correct, one would not expect to find the almost non-existent SVOM structure and there would be no need to invoke a new device which came to affect the epistemic pre-modals. This would avoid the complications noted above.)

(v) Another change which helped to isolate the pre-modals was the introduction of the to infinitive. The handbooks usually depict a titanic struggle between the to and to-less infinitives raging during the late OE and early ME period, with to eventually winning out in most places. The exact way in which the to infinitive came to prevail is a matter of some argument, but Visser reflects a common view in asserting that in the beginning the introduction of to was semantically based, that is, that it first appeared as a true preposition with a sense of 'direction towards' (which changed, for example, singan to singenne). Gradually the prepositional force of to became weakened and the use of the to infinitive increased rapidly during the early ME period. However, some ME glossaries and dictionaries alternate between the to and bare infinitives and Kellner (1892) points out that 'even in the sixteenth century the simple infinitive tries still to retain part of its old dominion'. The relevance for us is that the to form never occurs immediately after a pre-modal (although it often occurs when separated by other material: to do youre bidding ay we will, Townley Mysteries 266, or she tells me she'll wed the stranger knight, Or never more to view nor day nor night, Shakespeare, Pericles II v 17). As long as to was interpreted as a preposition, one could predict that it would never occur after the pre-modals, because they conveyed no notion of 'direction towards'. However, it is not clear to me why the pre-modals consistently resisted the encroachment of to after it had lost its prepositional force and was
taking over elsewhere. It seems that the pre-modals were already beginning to be identified as a unique class. The conditions for the invasion of to are complicated and mysterious. The to infinitive is rare in subject position before NE (to go is necessary), whereas the plain infinitive is very common in this position. On the other hand, the to infinitive in ‘extraposed’ positions (it is good to go) was introduced in OE and became common in ME (see §4.3). Any rate, Visser (1963–73: §901) concludes that ‘it took a long time for the particle to to be reduced from a preposition expressing motion, direction...to a semantically empty sandhi form, functioning as a mere sign of the infinitive...after about 1500 the construction with plain infinitive is on its way towards obsolescence’. So we must conclude that the pre-modals were seen as a unique class by the sixteenth century, if not earlier.

This, then, was the first stage of the story. Five independent changes took place which had the effect of isolating the pre-modals as a distinct class: (i) loss of all the direct object constructions with pre-modals, (ii) loss of all the preterite-presents except the pre-modals, thereby isolating the latter as a unique inflectional class, (iii) increased opacity of the past tense pre-modals might, could, should, would and must, (iv) special marking of epistemic pre-modals to avoid otherwise expected SVOM or it M[NP...], structures, (v) the development of to infinitives with almost all verbs except the pre-modals. In early OE the pre-modals had no characteristics peculiar to themselves, but by the end of the ME period they had become identifiable as a unique class by virtue of these five changes in various parts of the grammar. The changes did not occur simultaneously and were presumably independent of each other; at least, I can see no causal relationship between them. The pre-modals from earliest times shared another unusual but not unique property that the subject NP of the S they dominated was never phonologically realized in surface structure, i.e. was obligatorily PRO. Some other verbs, such as try, have this property.

(55)  
a. Bill could _ do that  
b. it may _ be that Bill left  
c. Bill tried _ to do that

2.2.3 These changes seem to have taken effect by the end of the fifteenth century: all direct objects had been dropped by then, except

1 Verbs of perception also preserve the plain infinitive: I heard him play ‘Bitches Brew’.
with can; the last surviving non-pre-modal preterite-present is purfan in c.1485; several new non-temporal senses of the preterites of the pre-modal came in before 1500; the Extraposition marking for epistemics developed with the SOV-to-SVO base change, in ME; Visser dates the final obsolescence of the plain infinitive from 1500. The pre-modal were now firmly established as a unique class inflectionally, syntactically and semantically. It was no longer so clear that they were true verbs; if they were, they had several exceptional features. Evidence suggests that at this time they were re-analysed as a new category, which we can call ‘modal’, being derived as part of an S containing the verb they govern; the grammar did not continue treating them as verbs and piling up more and more ‘exception features’. The evidence is a whole series of changes taking place simultaneously in the sixteenth century, all of which follow automatically from a re-analysis along the lines that I have suggested. It is remarkable that all of these changes took place together, within the same short period.

(i) The old pre-modal could no longer appear in infinitival constructions. The last instance attested by Visser is that appered at the fyrst to mowe stande the realm in great stede... 1533 More, Works 885 C1; and Thomas Cooper records the form to may in his Thesaurus Linguae Romanae et Britannicae of 1565.

(ii) The old pre-modal could no longer occur with -ing affixes. The last recorded cases I have found are the potential mode signifyth a thyn as mayying or owing to be done, c.1512 Linacre, Progymnasmata, and maeyinge suffer no more the loue and deathe of Aurelio, 1556 Aurelio and Isab. M ix.

(iii) As from the mid-sixteenth century there could be no more than one modal for any verb in the standard dialects. Visser’s last attested case of two modals is I fear that the emperor will depart thence, before my letters shall may come unto your grace’s hands, 1532 Cranmer, Letters. (In fact, some dialects of NE allow such sequences: Scots has will you can come? and southern US dialects allow she might could do it. I do not know if these are innovative or conservative forms.)

(iv) The old pre-modal could no longer occur with have and an -en affix. The last case seems to be if wee had mought conuenient come to-gyther, ye woulde rather have chosin to have harde my minde of mine owne mouthe, 1528 More, Works 107 H6.

These four changes follow automatically from deriving the pre-modal from M, introduced by a rule Aux → T (M). This entails that
there can be no more than one M in any S, that it cannot occur ad-
jectively in a participle form, that it will never be found in a for-to
complement type (since in that construction Aux is manifested only as
to) and that in any one S have and be will never appear on the left
of a modal. It is important to note the simultaneity of the changes.
Three further changes took place at about the same time, which,
while they are not automatic consequences of the re-analysis of the
pre-modals, do nonetheless follow in a natural way which I shall
explicate in §7.3.
(v) The Negative Placement rules had previously placed the negative
particle immediately after the first verb, whether or not it was a pre-
modal. This gave patterns like John could not take the bread and John
took not the bread. With the development of two categories, if the
existing rule were preserved, it would yield a new surface structure
John could take not the bread alongside John took not the bread. Under
this hypothesis, the grammar had to decide, as it were, whether the
negative should be attached to the right of the modal or of the real
verb, or whether to change the rule, generalizing to apply to both M
and V, thereby maintaining the existing surface structures. It took the
former course and retained forms like John could not take the bread.
For clauses containing just one verb, two patterns developed in the
early sixteenth century: he did not speak and, without do Support,
he not spoke. In both cases the evidence was that Negative Placement
had been re-formulated to place the negative between Aux and V.
Constructions like he not spoke began to decline sharply in the late
seventeenth century and the do type became standard. The conservative
I spoke not survives into modern times in isolated phrases such as
I think not, I hope not (although these structures represent I think
[not S], rather than I don’t think [S]), but it became much less common
from the late seventeenth century and was increasingly confined to
poetry and archaisms (see §2.3).
(vi) Parallel changes had to be made in the Subject-Verb Inversion
transformation, which had given patterns like could John take the
bread? and took John the bread? Now that could and took were analysed
as members of different categories, the inversion rule either had to be
generalized to operate on M and V, in order to maintain the old patterns,
or to prepose only auxiliaries.\(^1\) It took the latter course. Jespersen and

\(^1\) Logically it is possible that the grammar could have elected to invert only on the V,
but this would give new surface structures with very strange cross-over phenomena:
Visser show that did John take the bread?, where the verb is not inverted, becomes common in the sixteenth century and the verb-inverted took John the bread? is definitely on the decline in the latter half of the seventeenth century.

It is impossible to put precise dates on these last two changes since they are both inextricably intertwined with the development of the do Support rule, which was a lengthy and complicated process and not clearly understood by grammarians. Competing forms existed side by side for long periods. Do began to develop in ME, but Jespersen (MEG v 25.6) points out that 'the exuberant use of do, chiefly unstressed, and having no real grammatical value, was reached in the 16th c. . . . but from the 17th c. a reaction set in and gradually restricted the use of do to those three cases' (i.e. with negatives, inversions and to indicate emphasis). Hence these changes occurred at roughly the same time as the postulated re-analysis, but we shall return to the development of do Support in §2.3.

(vii) So the old pre-modals were no longer verbs and were no longer to be derived from higher predicates. But now came what is perhaps the most remarkable change of all. As this re-analysis took place, a new set of verbs came into the language. These are often called 'quasi-modals', namely the verbs be going to, have to, be able to. Despite some subtle nuances, these verbs are for most purposes semantically identical to the modals shall/will, must, can and differ only syntactically in that they have all the usual properties of other verbs. They are true verbs. It is as if the re-analysis of can, may, must, etc., as modals created a vacuum, which the grammar immediately filled by creating a new set of semantically equivalent verbs. Again the dates are striking. The first OED references for have to are he told him he had not to beleeve, that the couetousness of Virginio . . . had moved Ferdinand, 1579 Geoffrey Fenton, Guicciard (1618) 6, and we have . . . to strive with a number of heavy prejudices, 1594 Richard Hooker, Of the Laws of Ecclesiastical Polity i i I. The first OED reference for be going to is this unhapy soule . . . was goyng to be broughte into helle for the synne and onleful lustys of the body, 1482 Monk of Evesham (Arb.) 43. The situation with be able to is more complicated. The OED's earliest reference is to a felede uper lyeph leye . . . or is able to pasture . . ., 1398 Trevisa's
2.2 Historical re-analysis

Englishing of Bartholomaeus de Proprietatibus Rerum xiv p49, where able means 'suitable'. The earliest convincing example where to be able is equivalent to can, is to be abill to wed hure, 1440 The Gesta Romanorum, ed. S. J. H. Herrtage (1879), p269. As Strang (1970: p150) puts it, with the development of these forms 'English developed a repertoire of full verbal paradigms to supplement the defective and functionally restricted modals'.

2.2.4 One can summarize the re-structuring in the following (greatly simplified) fragments of grammars, where must, may, etc., are subsumed under V in OE, under M in ENE.

<table>
<thead>
<tr>
<th>Fragment of OE grammar</th>
<th>Fragment of ENE grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>S → NP Aux VP</td>
</tr>
<tr>
<td>VP → V (NP)</td>
<td>Aux → T (M)</td>
</tr>
<tr>
<td>NP → [N]</td>
<td>VP → (have + en)(be + ing) V (NP)</td>
</tr>
<tr>
<td></td>
<td>NP → [N]</td>
</tr>
<tr>
<td>Negative Placement</td>
<td>Inversion</td>
</tr>
<tr>
<td>NP V X ⇒ NP V neg X</td>
<td>NP Aux X ⇒ NP Aux neg X</td>
</tr>
<tr>
<td>Inversion</td>
<td></td>
</tr>
<tr>
<td>NP V X ⇒ V NP X</td>
<td>NP Aux X ⇒ Aux NP X</td>
</tr>
</tbody>
</table>

I have not 'proved' anything here and no doubt there are alternative ways of treating this data; but a radical re-analysis of pre-modal of the kind suggested here accounts naturally for the simultaneity of the seven changes taking place in the sixteenth century, and, as we shall see, it also permits an explanatory account of the historical changes. The relevant dates are remarkably uniform and 'clean' and there is no need to idealize the data, in a way that one would ordinarily expect to have to do, abstracting away from different dialects, literary styles, etc. The re-analysis is a genuine syntactic change affecting the base rules of the grammar of English. It is hard to see how the flush of changes taking place at the same time could be treated as anything other than an incredible accident in a more abstract, Ross-type analysis of the modern modals. One assumes that such an analysis would entail the claim that seven exception features were added simultaneously to various parts of the grammar. The historical corollary of abstract analyses of this type is that the base rules do not change and that radical re-structurings like this just do not happen (Lakoff 1968).
On the contrary, I have assumed that a grammar is a system of rules, not a fundamentally taxonomic system with a small number of rules smothered by large sets of rules features; such a view of grammar permits a description of the simultaneity of the surface changes discussed. These rules interlock in complex ways and changing one part of the system will cause 'pressure' on another part. Changes in the membership of an inflectional class, combined with some independent changes elsewhere in the grammar, were followed by the re-analysis of some verbs as a new category (manifesting itself in the changes (i)–(iv), (p)110), and that in turn precedes changes in certain transformational rules (e.g. Negative Placement and Subject-Verb Inversion) and the introduction of some new lexical items. The re-analysis was provoked by a number of changes which made it unclear whether the pre-modals were verbs or a unique category. Analysing them as a new category would avoid having to treat them as verbs with a cluster of exception features and would in a sense institutionalize their exceptionality. It seems that it did not take many exception features to bring about the re-analysis. Thus the category membership of pre-modals became opaque and the grammar moved to avoid such opacity. What this strongly suggests is that grammars are not as abstract as many transformationalists have been supposing over the last several years. It seems that some kind of transparency principle is at work in syntax and that initial structures are constructed and categories determined on the basis of fairly tangible and superficial alternations. As soon as may and, say, take started to behave differently with respect to things like inflections, infinitive forms and some other phenomena, there was a wholesale re-analysis of the pre-modals, keeping the initial structures within earshot of the surface structures; the grammar did not just go on acquiring more and more exception features. At least that hypothesis would explain the remarkable chronology of the changes I have discussed.

We shall return to the status of such an explanatory Transparency Principle in §3.1, and in §7.3 we shall specify a causal chain showing how the re-formulation of Negative Placement and Subject-Verb Inversion was a necessary consequence of the introduction of the modal category. The Transparency Principle, therefore, will be a principle of grammar which explains why this re-analysis took place. A theory of grammar, permitting only shallow derivations and incorporating the Transparency Principle, not only accounts for the
simultaneity of the various surface changes, as shown in §2.2.3, but also explains why this change should take place, as we shall see after developing a theory of change in chapter 3. Each of these feats is beyond the capacity of a more abstract analysis like that of Ross, or, it can be added, that of McCawley (1975). McCawley, echoed more recently by Pullum & Wilson (1977), argued that can, must, etc., are really verbs with a defective paradigm, lacking non-finite forms, i.e. participial and infinitival forms. This accounts for their inability to occur with -ing, -en, or as the (presumably infinitival) complement to another verb. However, it says nothing about why present and past participles and infinitives should be lost at the same time, why it should happen to these verbs and not to, say, take and go, why as these forms were lost there were at the same time changes in inversion and negative patterns, or why Tag Formation and Number Agreement distinguish these defective verbs from fully endowed verbs (Tag Formation applying to the defective verbs and not to full verbs, and vice versa for Number Agreement). Since this proposal cannot describe simultaneity of the changes, it certainly cannot give an explanation (see §3.1 for some synchronic defects of an almost identical proposal).¹

2.3 Appendix: periphrastic do

The aim of this chapter was to show that changes can take place in the base component of a grammar, in particular that a new category can be introduced, and that changes may affect only the syntactic component. So far I have argued that a modal category was introduced into the grammar of English, and that suffices to validate the general point. However, one may go on to ask whether this was the full extent of the change taking place in the sixteenth century. There is evidence that the change was more extensive and that the categories of Aux and T

¹ It is worth noting that such defective paradigms do not necessarily occur in non-finite packages. Stride, for example, lacks only the perfect participle: Susan is striding to victory, wants to stride to victory, *has stridden/strid(d)ed to victory.

We have argued for a historical change in the phrase structure rules of English. Under the usual assumptions, that entails a change in the base component. However, if somebody proposed a generative grammar without transformations relating different levels of representation (see §6.2 for a proposal to eliminate structure-preserving rules), one would still hold that there had been a change in the phrase structure rules, which under this view would generate surface structures without the intervention of transformations. Therefore, under certain relatively extravagant assumptions, the claim for an independent, syntactic change is independent of the claim that there was a change specifically in the base component.
were also introduced for the first time. The evidence concerns the periphrastic *do*.

Grammars of English, from *Syntactic structures* onwards, have analysed tense as a deep structure constituent, unlike grammars of, say, French, where tense is best treated as a feature on the first verbal element of a clause. In NE the tense marker has the same distribution as the auxiliary verbs, i.e. modals and aspectual markers: it occurs before the negative, before the subject in inverted questions, in anaphoric VPs, and in tag questions.

(56) a. John did not leave
    John must not leave
b. (who) did John leave?
    (who) must John leave?
c. Mary left and John did too
    Mary will leave and John must too
d. John left, didn’t he?
    John must leave, mustn’t he?

It is the availability of the periphrastic *do* which enables the tense marker to have this freedom of movement and independence of its verb. To adopt a transformational analysis, we may say that this was not possible before the development of the rule of *do* Support, which provides a ‘support’ for tense markers ‘stranded’ by independently motivated rules (i.e. Negative Placement, Subject-Auxiliary Inversion, VP Deletion and Tag Formation) and no longer situated to the immediate left of a verbal element. *Do Support* applies after Affix Hopping which attaches T to any verbal element to its immediate right; *do* Support operates on all unattached or stranded Ts.

\[
\text{id Support} \quad X \ T \ Y \overset{\text{oblig}}{=} X \ do + T \ Y
\]

At a time when there was no periphrastic *do*, the patterns of (56) did not occur and there would be no reason to treat tense as anything other than one of the features on the verb, like number and person. If T was not an initial structure constituent and if *may, must*, etc., were verbs, then there would also be no justification for the category Aux (recall that the aspectual markers are analysed as constituents of VP). It seems plausible to claim that *do* Support was introduced at the same time as the modal category, although the evidence is not as
clear-cut. If this is correct, the full extent of the rule changes affecting the auxiliary elements in the sixteenth century was a new phrase structure rule $\text{Aux} \rightarrow \text{T (M)}$ and a new transformation, $\text{do}$ Support (in addition to the re-formulations of Negative Placement, Subject-Auxiliary Inversion, and the new lexical items, $\text{have to, be going to, be able to}$). The evidence will be sketched only briefly, since I do not intend to take a position on the nature of this change.

Some have claimed that the periphrastic $\text{do}$ dates from OE or even earlier. There is a similar construction in certain dialects of German, $\text{tun} + \text{infinitive}$ (Erben 1969), but this must be viewed as an independent development. $\text{Do} + \text{infinitive}$ in OE was always a causative and quite unlike the modern usage. Similarly German $\text{tun}$ occurred frequently in early New High German literary texts of the fourteenth-sixteenth centuries, but was strictly a causative. $\text{Tun}$ and $\text{do}$ underwent parallel but presumably independent development and eventually took on similar tense-carrying functions, although $\text{tun}$ is now vanishingly rare in the standard language. This parallelism does not indicate an ancient ancestry, but rather a common innovation.

In his monumental study Ellegård (1953) argues that the non-causative, periphrastic $\text{do}$ does not come into existence until the late thirteenth or early fourteenth century. He claims that the change of use took place under the special conditions of the southwestern dialect in the thirteenth century. ‘The decisive factor was that causative $\text{do}$ in this dialect had come to be used almost solely in positions where a periphrastic interpretation was also possible without changing the meaning of the whole context. This ‘equivocal’ $\text{do}$ had an exact correspondence in French, which undoubtedly had some influence in the spread of its use’ (p208). The equivocal case arises where the infinitive has no subject: $\text{he dude writes sende}$ ‘he caused to send writs’, $\text{fae gert seke him in pat sesoune}$ ‘they caused to seek him in that season’ (ger is the Scandinavian counterpart of English $\text{do}$). Out of context such constructions may be causative or periphrastic, such as $[\text{he}] \text{dude strepe bis Maide naked}$, which might mean that he had this maid stripped or that he stripped her himself. Infinitives with subjects can only be causatives: $\text{fae gert hym bere it with envy}$ ‘they made him bear it with envy’. Actual equivocal cases (i.e. cases not disambiguated by the context) were particularly common in texts of western dialects from the thirteenth century. Ellegård showed that the periphrastic $\text{do}$ was exploited first by poets for rhythmical purposes and rhyming (since its
availability often permitted the infinitive to occur at the end of a line), pointing to some eastern writers who used the periphrastic in verse but only the causative in prose. As the periphrastic became more common, so the causative gave way to \textit{make} and \textit{cause}. 'It is not until the end of the fifteenth century that the do-form becomes widely used in prose texts. From then on it spreads fast for about two generations' (p209).

As noted earlier, \textit{do} was used exuberantly at this time, particularly in learned writings, occurring even with pre-modals: \textit{hit you behowith... behold ho [= who] shall doo gouterne}, c.1475 Partenay (EETS) 2385. This usage survived until St Thomas More: \textit{now if I would then doe... tel him that...}, 1534 Works (1557) 1192 F4. Sweet (1900: p88) says of this period that the do-pattern was used 'according as a caprice, convenience, and clearness of construction, or euphony suggested', although there were constraints, e.g. \textit{do} did not occur before be or have in 'standard' English (Irish dialects use it even here: \textit{would anyone believe the things they do be saying in the glen?}, Synge, Shadows of the Glen).\footnote{The first example of the periphrastic \textit{do} in imperatives is \textit{come, do be a good girl}, \textit{Sophy}, 1749 Fielding, Tom Jones (Everyman) vol. II p286.}

From the fifteenth century the use of \textit{do} was more strictly confined to its present-day environments, and this state was reached finally in the seventeenth century, although as the periphrastic \textit{do} originated in southwestern dialects, so according to Palmer (1965: p26) those dialects have preserved the 'exuberant' uses more tenaciously. Contemporary grammarians viewed \textit{do} as a mere tense carrier, semantically empty. Ch. Butler in his \textit{English Grammar} of 1633 (p45) says 'The Present tense is formed of the Right case [i.e. the infinitive], either with the sign, [i.e. auxiliary \textit{do}], or without the sign: 'I loove, thou loovest, hee looveth...'; with the sign: \textit{I doo loove, thou doost loove, hee dooeth loove...} – The Imperfect is formed either of the oblique Case [i.e. preterite] without the sign, or of the Right case with the sign, thus: \textit{I loooved or did loove}'. C. Cooper agrees in his \textit{Gramm. Ling. Anglic.} (Pt III Cop4 §5) of 1685: '\textit{I did learn discebam et I learned aequipollent}'.

In the light of this, it is plausible to claim that \textit{do} Support became firmly established in the grammar of English by the seventeenth century. As noted in §2.2.3, the rule was established at the same time as Negative Placement and Subject–Auxiliary Inversion were reformulated to accommodate the re-analysis of the pre-modals. Before the re-analysis the canonical surface structure patterns included those
of (57) for negatives and interrogatives, where V encompassed full verbs and pre-modals, and always carried the tense marker.

(57)  
  a. NP  V not ...
  b. V  NP ...

After the re-analysis the canonical patterns were (58), where Aux included modals and do, and carried the tense marker.

(58)  
  a. NP  Aux not ...
  b. Aux  NP ...

It was possible to assign a fairly precise date to the re-analysis of the pre-modals; this is more difficult for the introduction of do Support. There is clear evidence for a do Support transformation when the periphrastic do becomes limited to interrogatives, negatives, anaphoric VPs and tag questions.¹ There is no evidence for such a rule when do occurred only as a causative, when it would be treated as a normal lexical item occurring in the initial phrase marker. What about the intervening period? At this time do appears as an optional tense carrier and again, therefore, there is no reason to postulate a do Support transformation. However, T does occur as a constituent independent of the main verb of the clause. A grammar could handle this in a variety of ways, including a phrase structure rule Aux → T or treating do as a 'higher predicate', a true complement-taking verb. I know of no compelling evidence for a choice between these two options; each adds complications to the grammar. Therefore, while arguing that the present version of do Support was introduced just after the re-analysis of the pre-modals, I cannot take a position on the introduction of Aux and T as initial structure constituents: they were

¹ If we look at this as the development of a do Support transformational rule and T as an initial structure constituent, then it is an unexplained fact that the auxiliary do occurs first in anaphoric VPs, as in (i).

(i)  
  a. reced weardode unrim eorla, swa he oft ær dydon ‘innumerable men guarded the building, as they often did’. Beowulf 1238
  b. he her curseth as other fooles do. c.1503 More, Book of Fortune (Works 1557) 3 D2.

These expressions occur freely in the very earliest texts, and are the only evidence for treating T as a constituent in OE. Perhaps this evidence was sufficient to force the analysis. If so, one might argue that English had a phrase structure rule Aux → T from the earliest times and that do was attached first only to a T where there was no verb in the clause, then to any T (except before be or have), and finally only to a T not immediately preceding its verb. This would account for the relevant chronological stages, whereby do occurred (1) only in anaphoric VPs, (2) almost anywhere, and (3) in inverted interrogatives, negatives, anaphoric VPs and tag questions.
certainly introduced by the sixteenth century, but may have been introduced before the re-analysis of the pre-modals, when the periphrastic do became widely used as an optional tense carrier in the prose texts at the end of the fifteenth century. In either case one can say (if one rejects the hypothesis in p119n) that Aux was introduced as a category within the period of 1475–1550. This is roughly simultaneous with the introduction of the modal category.