SYNTACTIC CHANGE

A MINIMALIST APPROACH TO GRAMMATICALIZATION

IAN ROBERTS
Downing College, University of Cambridge

AND

ANNA ROUSSOU
University of Patras

CAMBRIDGE
UNIVERSITY PRESS

themes of this book is that the right kind of formal approach to grammaticalization can be quite revealing. We offer our ideas as an attempt to shed light on a central and intriguing property of language, which is clearly of common interest, from a novel perspective. We hope that this book will also be relevant to those researchers who may not be interested in issues of syntactic change, but are interested in theoretical questions such as the notion of functional categories and the nature of parametric variation.

1 *Parameters, functional heads and language change*

1.1 Introduction: the logical problem of language change

In the Principles and Parameters framework cross-linguistic variation is accounted for by means of assigning different values to a finite set of options, called parameters, that are provided by Universal Grammar (UG). In Chomsky (1981, 1986a) parametric options are associated with the principles of UG. To take an example, consider the Extended Projection Principle (EPP), which basically requires that all clauses have a subject. A parameter then determines whether this subject, when pronominal, is always overtly realised (in finite contexts at least). It is in English; it does not have to be in Italian. This is the 'pro-drop' – or null-subject – parameter; its effects are illustrated with the Italian and English examples in (1a) and (1b) respectively:

(1) a. Parla italiano.
'He/she speaks Italian.'
b. *Speaks Italian

In this model, the task of language acquirers is to set the right parametric values on the basis of the input they are exposed to. Thus UG along with the appropriate trigger experience yields a particular grammar. The task of the linguist, on the other hand, is first to identify the UG principles, and second to define the class of associated parameters. It is clear that the simplest possibility is that parameters are restricted to just two values; this desideratum has been largely followed in the literature.

Although this approach to parameterization seems to work for cases like the 'pro-drop' parameter in (1), it turns out to be insufficient once a wider range of parameters is taken into account. Consider, for example, Binding Theory, and in particular Binding Principle A, which states that an anaphor must be bound in its local domain. As Wexler and Manzini (1987) show, the notion of the local domain can be defined as the category that contains the anaphor and one of the following: (i) a subject, (ii) Inflection, (iii) Tense, (iv), indicative Tense, or finally (v) a root Tense. In other words, Binding Principle A is

subject to a five-valued parameter. Moreover, it is possible to find languages that make use of more than one value, depending on the type of anaphors they possess. Dutch is an example, as it has two types of reflexives, namely *zich* and *zichzelf*, which have distinct distributional properties. In particular, *zich* accepts a long-distance antecedent, while *zichzelf* behaves more like the English reflexive *himself/herself*, thus requiring a local antecedent (parameter (a) in the Wexler and Manzini (1987) system). This is illustrated in (2a) and (2b) respectively (cf. Koster and Reuland 1991 for an overview of the data):

(2) a. Max_i bewondert zichzelf$_i$/*zich$_i$.
'Max_i admires himself$_i$.'
b. Jan_i liet mij voor zich$_i$/zichzelf$_i$ werken.
John made me for him work
'John_i made me work for him$_i$/*himself$_i$.'

Wexler and Manzini (1987) concluded that parameters must be associated with lexical items, offering further support for Borer's (1984) original claim. Regarding (2) then, the choice of the antecedent is a lexical property of the elements *zich* and *zichzelf*, and as Pica (1987) showed, it correlates with the internal structure of the reflexives. Attributing the parameter to the lexical properties of the anaphors allows us to maintain Binding Principle A as a non-parameterized principle, which states that anaphors must be locally bound. Parametric variation with respect to what counts as local is associated with the relevant lexical items.

The idea that parameterization is restricted to the lexicon has been successfully pursued in subsequent research, which has further limited the set of parameterized lexical items to functional categories (see Chomsky (1995, 2000) for a recent statement). Language acquisition is still seen as the process of parameter setting, albeit as specifically fixing the values associated with functional categories. It is uncontroversial that the lexicon has to be learned, and, on this view, parameter setting reduces to a facet of lexical learning. We can now view the initial state of UG as consisting of a number of principles and of open parametric options; the latter are associated with a specific set of lexical items, the functional categories. To illustrate this, let us reconsider the 'pro-drop' parameter: the EPP is not parameterized, but the inflectional category responsible for subject agreement, call it AgrS, is. In particular, if AgrS is in some sense rich enough, that is, has the right properties, to license and identify an empty pronominal subject, we have the Italian setting, yielding (1a); if not, then we have the English setting, predicting the ungrammaticality of

(1b) (Rizzi 1986a). Roughly speaking, children have to determine, on the basis of experience, whether their language has the English-type or the Italian-type AgrS. Similar considerations extend to (2). Questions that remain open in current research include the characterization of the class of functional heads and the definition of the ways in which functional categories can be parameterized. The latter point is discussed in section 3 of the present chapter. We consider the former question in more detail in Chapter 5, where we will make some suggestions based on the evidence provided by grammaticalization.

According to what we have said so far, the acquisition of syntax is viewed as the process of parameter setting. Within this framework of assumptions, syntactic change can be viewed as change in the parametric values specified for a given language. In other words, parameter values can change as a function of time. We can in fact observe this very easily by comparing the Modern Romance languages with Latin in respect of word order. Latin word order was rather free, but object–verb order clearly predominated; on the other hand, the Modern Romance languages are all verb–object. The contrast is illustrated in (3), with Italian representing Modern Romance:

(3) a. Ego . . . apros tres et quidem pulcherrimos cepi. (Pliny the Younger)
(Object) (Verb)
I boars three and indeed very-beautiful have-taken.
b. Io . . . ho preso dei cinghiali, tre e anche bellissimi.
(Verb) (Object)
'I have taken three and indeed very beautiful boars.'

Thus, if there is a parameter determining the relative order of verb and direct object, its value has changed in the development of Latin into Romance. The central issue for diachronic syntax in the context of Principles and Parameters theory is accounting for how and why this can happen.

Following a view that has been developed in terms of recent linguistic theory, primarily by Lightfoot (1979, 1991, 1998), we assume that parameter change is an aspect of the process of parameter setting. A change is initiated when (a population of) learners converge on a grammatical system which differs in at least one parameter value from the system internalized by the speakers whose linguistic behaviour provides the input to the learners. As the younger generation replaces the older one, the change is carried through the speech community. Of course, many social, historical and cultural factors influence speech communities, and hence the transmission of changes (see Labov 1972, 1994). From the perspective of linguistic theory, though, we abstract away from these factors and attempt, as far the historical record permits, to focus on change purely as a relation between grammatical systems.

The assumption that parameter change is an aspect of the process of parameter fixation raises an important issue for language acquisition. The issue is summed up in the following quotation from Niyogi and Berwick (1995):

> it is generally assumed that children acquire their . . . target . . . grammars without error. However, if this were always true, . . . grammatical changes within a population would seemingly never occur, since generation after generation children would have successfully acquired the grammar of their parents. (Niyogi & Berwick 1995:1)

As the above quotation shows, the standard paradigm for language acquisition is not immediately compatible with the observation that grammatical systems change over time. To be more precise, it is generally assumed that language acquisition is a deterministic process: its final state converges with the target grammar that acquirers are exposed to. However, if convergence is always guaranteed, then the crucial question is how changes can ever take place. Clark and Roberts (1993, 1994) refer to this issue as the logical problem of language change, and sum it up as follows:

> if the trigger experience of one generation, say g_1, permits members of g_1 to set parameter p_k to value v_i, why is the trigger experience produced by g_1 insufficient to cause the next generation to set p_k to v_i? (Clark & Roberts 1994:12)

The simple answer to this question, which again goes back to Lightfoot (1979), is that v_i is unlearnable. In this case language acquirers have to revert to some other parametric option, thus triggering a change in the system. This way, the new setting for parameter p_k amounts to parameter resetting in comparison with the target grammar. If this is correct, we have to weaken and refine the notion of determinism, along the following lines: language acquisition is deterministic to the extent that all parameters have to be set. This allows for p_k to receive a different value from that found in the input, therefore making space for language change. This of course does not imply that changes have to take place; indeed, most of the time convergence is 'successful' in that children arrive at the same parameter values as their parents – this is reflected by Keenan's (1996) principle of inertia (see also Longobardi 2001a). A change occurs when the trigger experience for a parameter setting provided by the input has become obscure or ambiguous. This can happen in a variety of ways, for example through language contact, morphophonological erosion, etc. Fleshing this idea out requires us to develop an account of the relation between the learner and the trigger; it also requires us to be very precise about the nature and format of parameters. We will discuss parameterization in section 1.3; here we will focus on the relation between the learner and the trigger.

The logical problem of language change interacts with the logical problem of language acquisition. For the latter, the question is how children succeed in setting the parameters correctly on the basis of the input they receive, given that this input may be insufficient and degenerate (see the 'poverty of stimulus' argument of Chomsky 1986a). If by 'correctly' we mean complete matching with the adult setting, then the logical problems of language acquisition and language change become contradictory. If, however, by 'correctly' we mean simply fixing a value consistent with the trigger experience, as suggested above, then the contradiction does not arise. Let us call this the weakly deterministic view of language acquisition: the goal of acquisition is to fix parameter values on the basis of experience – all parameter values must be fixed, but there is no requirement for convergence with the adult grammar (although this happens most of the time).

The relationship between the learner and the trigger can be thought of as mediated by a device which takes experience as input and produces parameter values as output. The trigger experience is naturally thought of as consisting of sets of sentences (cf. Clark & Roberts 1993, Gibson & Wexler 1994, among others). Lightfoot (1998) and Dresher (1999) argue that learners use input forms as 'cues' for setting parameters. The trigger in this case is not sets of sentences but fragments of utterances (partial structures) (cf. also Fodor 1998). For Dresher (1999) each parameter has a marked and a default setting, and comes with its cue, as part of the UG specification of parameters. Lightfoot (1998:149), however, takes a much stronger view and argues that 'there are no independent "parameters"; rather, some cues are found in all grammars, and some are found only in certain grammars, the latter constituting the points of variation'. Let us illustrate this with the loss of the verb-second (V2) phenomenon in Middle English. The presence of exactly one constituent other than the subject in immediately preverbal position is a cue for the learner that a given language is V2. According to Lightfoot (who follows Kroch & Taylor 1997), the Northern dialects of Middle English had a V2 grammar, which at some point ceased to exist – Modern English is not V2, as the grammaticality of sequences like *Yesterday John left* shows. Lightfoot proposes that the change was triggered by the following: (a) interaction with speakers of Southern dialects which didn't have obligatory V2 and also didn't treat subject pronouns as clitics, so the *XP–subject pronoun–V* sequence in the input was evidence against a positive setting for the 'V2 parameter'; (b) the independent loss of all verb movement operations,

making verb movement to the second position impossible, pre-empting many V2 orders. In this way, the occurrence of the V2 cue was considerably reduced, leading to the consequent loss of V2. This approach, however, seems to involve circularity. It appears that V2 was lost because it was not cued, and that the cue was lost because V2 was undermined (owing to the factors given). It is not clear what the notion of cue is really explaining here; if we omitted it from our account, we would nevertheless have at least a plausible description of how V2 was lost. Also, Lightfoot's approach seems to involve a category mistake: cues are fragments of the trigger experience, sequences such as *XP–V* in the case of V2. But parameters are abstract properties of grammars, features of part of an individual's mental representation (his/her I-language). Although the notion of cue is useful, it must be kept distinct from the notion of parameter. Finally, Lightfoot's approach is too unconstrained: if there is no independent definition of cues, then we have no way of specifying the class of possible parameters, and hence the range along which languages may differ (synchronically or diachronically).

It is, however, possible to maintain that parameters can be independently defined and that learners also make use of cues provided by the input (this is closer to Dresher's view). Recall that according to current assumptions in the Principles and Parameters framework, parameters are lexical; it is also generally accepted that the lexicon has to be learnt, as it is language specific. There must be some learning device that enables acquirers to learn words (their syntactic, morphological, phonological and semantic properties). If parameters are linked to a subclass of lexical items, that is, functional elements, which also have to be learnt, then it follows that the same device is also responsible for setting parameters. This device may be part of UG, or it may be a separate device which interfaces with UG (we will tentatively assume the latter, mainly for clarity of exposition). Any part of the input that can provide the acquirers with information about the lexicon is a cue. This approach, unlike Lightfoot's, allows us to maintain both the notions of cues and parameters: cues are provided by the input, parameters are specified by UG and are set by the learning device on the basis of the interaction of cues and UG. The relation between the cues and the parameter values is indirect and is mediated by the learner.

We can make the notion of cue clearer if we consider the notion of parameter expression introduced by Clark and Roberts (1993:317):

(4) Parameter expression:
A sentence S expresses a parameter p_i just in case a grammar must have p_i set to a definite value in order to assign a well-formed representation to S.

As Clark and Roberts (*ibid.*) say: 'When a given datum expresses some parameter value, the learner will be under pressure to set that parameter to the value expressed by the datum.' This given datum is the trigger and is defined as in (5):

(5) Trigger:
A *sentence* S is a trigger for parameter p_j if S expresses p_j.

We can relate this notion of trigger to the notion of a cue by replacing 'sentence' in (4) and (5) by 'substring of the input text', as follows:

(4′) Parameter expression:
A substring of the input text S expresses a parameter p_i just in case a grammar must have p_i set to a definite value in order to assign a well-formed representation to S.

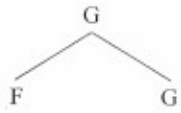
(5′) Trigger:
A substring of the input text S is a trigger for parameter p_j if S expresses p_j.

It is an empirical question what a substring may be. Arbitrarily, we will suppose that a substring can be no smaller than a morpheme (we are thus proposing that a morpheme is the minimal unit of grammatical analysis for language acquirers as well as for linguists) and no larger than a sentence (cf. Fodor 1998:17 for a similar proposal). If the parameter expression is robust enough, it will lead to the correct parameter setting. If, however, the parameter expression is ambiguous, then there must be some 'safety mechanism' in the learning device which leads to the assignment of a value – weak determinism requires this. This value will still be compatible with the input, but – again due to weak determinism – may differ from that of the target grammar, thus yielding a change.

The questions now are: (a) what is 'robust' parameter expression? (b) what is the 'safety mechanism' referred to in the above paragraph? We know of no good general answer to question (a), beyond observing that many parameters seem to be morphologically expressed, and when independent morphological or phonological changes conspire to remove or obscure this expression, a parameter change may take place (see Roberts 1999 on this). This answer is undoubtedly insufficiently precise and insufficiently general. Concerning question (b), we suppose, following Clark and Roberts (1993), that the learning device is computationally conservative in that it has a built-in preference for relatively simple representations. In other words, if the trigger is ambiguous, the learner will choose the option that yields the simpler representation. We will consider the question of how to define simplicity in detail in Chapter 5, but here we will provide a preliminary illustration of what we have in mind. Let

us assume that movement operations are adjunctions, as proposed by Kayne (1994); then movement always creates relatively complex representations, in the obvious sense that (6b) with F adjoined to G is a more complex structure than (6a), where no movement, and thus no adjunction, has taken place:

(6) a. G

b. 

(Here G and F may have any amount of internal structure; in particular they may be either heads or XPs.) Loss of movement will lead to a reduction in complexity, that is, to a simpler representation. More precisely, if the learner postulates non-movement the simplicity preference will be satisfied. So movement must be robustly triggered (note that we are implicitly taking movement to be a parameter here – we develop this in section 3). If (6b) is not properly triggered, then (6a) will be preferred. Where (6b) changes to (6a) a movement operation is lost. However, there is another possible outcome where (6b) changes to (6a). The learner may analyse some instances of the moved category F as part of the inflectional system instantiated by G (this idea depends on the assumption that movement is always and only to a functional position – see section 3). This kind of 'misanalysis' results in recategorising a class of lexical elements as inflectional items; in (6b) F is reanalysed as G, essentially giving the structure in (6a). In other words, 'misanalysis', in the sense described here, can create new functional material. We will argue extensively in Chapters 2 to 4 that this kind of structural simplification is precisely the one that occurs in cases of grammaticalization. Another kind of structural simplification involves reanalysis of an XP, a category with a certain amount of internal syntactic structure, as a simple head X, a category with no internal syntactic structure. The same considerations relating to language acquisition apply to this kind of reanalysis as to the loss of movement, and we will see in Chapters 2 to 4 that this kind of reanalysis, among others, is also prevalent in grammaticalization.

To summarize, in this section we considered the general assumption that parameters are a property of lexical items. We discussed the general approach to language acquisition in the Principles and Parameters model, according to which the process is viewed as parameter setting. Syntactic changes, on the other hand, are the result of changing parametric settings. Learnability issues

connect to both language acquisition and language change, as there has to be some mechanism that allows the learner to set or reset parameters on the basis of the trigger experience. The latter happens when the trigger (or cue) is obscure. In this case, we propose that the learner will opt for the default option as part of the built-in preference of the learning device for simpler representations. The logical problem of language change is addressed in terms of the idea that the learning device is computationally conservative; a value v_i of parameter p_k can be changed where the trigger experience (or cue) for v_i is not sufficient to prevent a simpler option being chosen. This 'insufficiency' of the trigger experience can arise through the effects of other syntactic changes, phonological changes, language or dialect contact, etc.

One question that this approach gives rise to is: why are grammars not tending towards some maximally simple state, which, at the very least, would be free of movement operations? The answer is that the simplifications effected by changes are always local, and may increase complexity elsewhere in the system. In fact, grammaticalization is a case in point: as already noted by Meillet (1912), grammaticalization may increase the notional expressive power of the grammatical system. Von Fintel (1995:184) notes that under grammaticalization 'the meaning of a lexical category is composed with a functional meaning to yield a new, more complex functional meaning'. In our terms, grammaticalization may provide a functional category with new exponents – this will become clearer in the next two sections. But, as just sketched, grammaticalization nevertheless arises from the learning device's bias towards simpler representations.

In the remainder of this chapter we will make more precise what it means to say that parameters are lexically associated with functional categories, beginning with a general discussion of functional heads themselves.

1.2 Functional categories

In the previous section, we presented the recent Principles and Parameters approach to cross-linguistic variation, according to which parameters are associated with functional categories. Parameterization as such then is restricted to the lexicon. Syntax connects the Phonological and Logical Forms (PF and LF respectively), that is sound and meaning. This is achieved with the help of the two basic mechanisms: Merge and Attract/Agree (Chomsky 1995, 2000). Merge is a binary operation that recursively combines elements, thereby building phrase structure. Agree is the operation that manipulates combinations, by establishing a relation between lexical items within a syntactic space. A simple

example is the agreement that we see between the subject and the verb in a sentence like:

(7) John likes/*like apples.

The agreement *-s* on the verb is the morphological expression of the relation that holds between the subject and the verb.

Lexical items belong to various categories, and this information is relevant for the syntactic operations of Merge and Agree. Categories then are primitive symbols associated with lexical items. The distinction between lexical and functional categories has its antecedents in traditional grammar. While Nouns, Verbs, Adjectives, and (at least some) Prepositions are lexical categories, elements such as Tense, Complementizers, Determiners, Negation, to name but a few, belong to the set of functional categories. The distinction between two kinds of item is an old one (cf. the Aristotelian distinction between *substance* and *accidence*), and it comes under various names, such as open versus closed class or lexical versus grammatical categories. In general, the basic distinguishing property is that lexical categories have descriptive content while functional categories do not; instead they carry grammatical meaning (cf. Radford 1997, Chapter 2 for a recent introductory discussion).

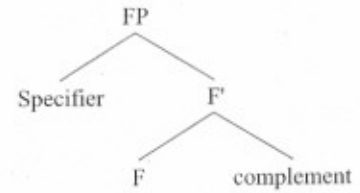
This distinction is widely accepted as one which holds in the lexicon. The question though is whether functional categories also have a syntactic representation. In other words, are functional categories also subject to syntactic operations, such as Merge and Agree? Some grammatical material seems to be purely morphological, and to have no role to play in syntax; for example, this seems to be the case of noun declensions or verb conjugations in languages like Latin or Classical Greek. Other material appears simply to duplicate other elements in the clause. For example, negation in French is realized by means of two elements, that is, *ne . . . pas* as in (8a), although only one of them (*pas*) is the 'true' negation. Similarly, expletives such as *there* in English double the postverbal subject in a construction like (8b) (the same can be argued for the subject agreement that we see on the verb in (7)):

(8) a. Marie *n'*aime *pas* Jean.
Mary not loves not John
'Mary doesn't love John.'
b. *There* arrived *three students*.

We can also see from (7) and (8) that grammatical material is lexically specified for morphological properties: *pas* in (8a) is a free morpheme, while *-s* in (7) is a bound one. Most importantly, grammatical properties such as those in (7)

and (8) turn out to be relevant in syntax as well: subject–verb agreement as in (7) is sensitive to the syntactic notion 'subject', while both *ne* and *there* in (8) have been argued to be syntactic markers of different kinds of scope (see Kayne 1984 on *ne*, Williams 1984 on *there*), also a notion standardly defined over syntactic structures. If this is correct, then we have to ensure that they are somehow syntactically present. In other words, we need syntax to be able to make reference to features associated with functional categories. Now, since Chomsky (1970), categories have been analysed as feature matrices. This means that it is possible – and, given the considerations just raised in connection with (7) and (8), desirable – to analyse grammatical features like agreement, negation, tense, etc., as syntactic categories. Given both the standard view of phrase structure (X'-theory), and the more recent Bare Phrase Structure of Chomsky (1995), that means grammatical features can function as heads which project a phrasal category containing a specifier and a complement, as follows (cf. (2) of the Introduction; here F is any feature):

(9)



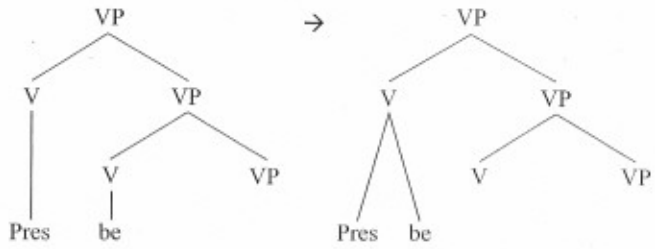
We will now provide some arguments in favour of having functional elements syntactically present in the sense just described. Let us begin by looking at the English auxiliary system:

(10) a. Do you like fish?/ *Like you fish?
b. I don't like fish/ *I like not fish.
c. Fred likes fish and Bill does/ *likes too.
d. I should go/ *I should do go/ *I do to go.

As the examples in (10a)–(10b) show, main verbs do not invert and cannot support negation; in both cases the auxiliary *do* must be present. Similarly, *do* can occur in elliptical contexts, while main verbs cannot – see (10c). The examples in (10d) show that the auxiliary *do* is in complementary distribution with modals, such as *should*, and with the infinitival marker *to*. The empirical evidence in (10) is the standard way of distinguishing between auxiliaries and main verbs. In the late 1960s and early 1970s, there was some debate as to whether data like that in (10) justified the postulation of a distinct category Aux, or whether it simply meant that certain verbs (e.g. *do, shall, have, be*, etc.)

were to be assigned various kinds of exception features. McCawley (1971) took the latter view. As he points out: 'Auxiliaries are exceptional by virtue of undergoing a transformation "tense-attraction", which combines them with the immediately preceding tense morpheme. All other transformations that might appear to treat auxiliaries in a special way (for example, subject verb inversion) are simply transformations that follow "tense-attraction" and have a structural description calling for the first verb.' He proposes the following structure:

(11)



In (11) the auxiliary *be* is attracted by the higher V which carries tense information (i.e. Present in this case). A similar approach is taken by Emonds (1970, 1976) who argues that verb raising attaches *have* and *be* (which, unlike modals, are treated as members of V) to the Aux node. Verb raising next feeds subject inversion. *Do*-insertion inserts *do* under V to the left of the main V, while it is verb raising again that places *do* under Aux. Finally, *do*-deletion deletes *do* where Aux appears immediately adjacent to VP. What Emonds and McCawley have in common is that they isolate a given position – the highest V for McCawley, Aux for Emonds – as the structural position associated with tense-marking, and that auxiliary verbs can move into that position. This is the position Chomsky (1981) called I(nflection), and which more recently has become known as Tense.

In addition to noting the common points between McCawley and Emonds, we can make two further observations about the English auxiliaries, both of which are relevant for understanding the notion of functional category. First, even if we categorize auxiliaries *do, be* and *have*, the modals, tense, even the infinitival marker *to* (see Pullum & Wilson 1977) as verbs, we have to accept that they are morphologically irregular, have special syntactic properties and form a closed class of items. It is also important to observe that they lack a central lexico-semantic property of verbs, namely argument structure (with the possible exception of dynamic modals; see 2.1). Second, tense, modals and auxiliaries project like other categories. In current terms, this means that Tense heads the phrasal category TP. Its Specifier is arguably the subject position

(this is proposed in Chomsky (1995, 4.10)) and its complement may be VP. Consider (10a–b) again: subject–Aux inversion indicates that *do* must have a syntactic position, as it can invert with the subject; the same holds for (10b) as *do* can support the negative element *not*. Furthermore, as Ross (1967) argued, deletion processes show that auxiliary elements are part of syntactic phrases:

(12) Fred could have been killed and Bill (could (have (been (killed)))) too.

As (12) indicates, any of the bracketed material can be deleted. Ross argues that the simplest account of these facts is to treat each bracketed constituent as a separate VP, headed by the respective verb or auxiliary. If we treat the highest auxiliary (*could* in (12)) as Tense, then we have a reason for thinking that it forms a constituent with the following bracketed material (this constituent may be T′, if the Specifier of TP is the subject, as just mentioned; this point does not alter the fact that deletion processes show that functional heads project phrasal categories).

Having seen some evidence in favour of TP, let us now turn to another functional element, the Complementizer. Bresnan (1972) argued for the syntactic presence of a C(omplementizer) position as part of the extended structure of the sentence notated as S′ (as distinct from the 'core' sentence S, so subordinate clauses were assigned a structure like *[$_{S'}$ that [$_{S}$ John left]]*). Clearly, Complementizers like *that* and *if/whether* differ in that the former appears with declaratives, while the latter introduce embedded interrogatives. The C position thus carries information about clause-type, and as such it is natural to think of it as the head of the subordinate clause. Given that C can also bear the +wh specification (as in *whether*-clauses), its Specifier can be identified as the landing site of wh-fronting (see Chomsky 1986b). The structural complement of C is TP.

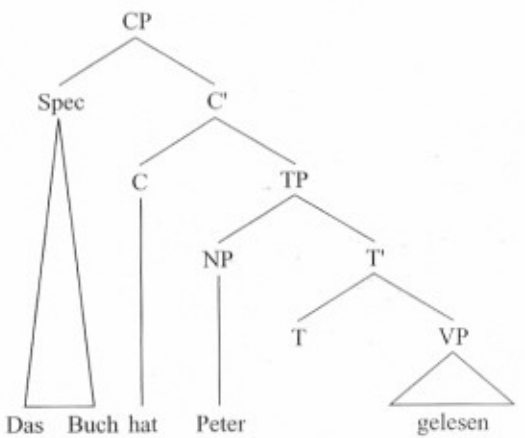
C plays an important role in the analysis of other phenomena. For example, den Besten (1983) showed that many main-clause inversion processes target this position, so C must also be the head of main (or root) clauses:

(13) a. Peter hat das Buch gelesen.
Peter has the book read
b. Das Buch hat Peter gelesen.
*Das Buch Peter hat gelesen.
'Peter has read the book.'

This is the well-known case of the verb-second (V2) phenomenon found in root declaratives in nearly all Germanic languages (and already mentioned in the previous section). The obligatory subject–Aux inversion in (13b) suggests

that the auxiliary has moved to a higher position, namely C. Notice that even if we treat auxiliaries as a class of defective verbs, we cannot do the same with complementizers. Thus, C at least is a grammatical element that has syntactic reality. The availability of inversion in matrix declaratives in V2 clauses, as in (13), brought to attention the intimate relation that holds between complementizers and auxiliaries, or to be more precise the relation between the C and T heads (cf. Platzack 1987 on Germanic, tensed complementizers in Irish discussed by Cottell 1995, etc.). This relation of course further supports the claim that T elements must be syntactically represented. Thus there are at least two functional elements, C and T, that project syntactically. This gives rise to a structure like the following for an example such as the well-formed sentence in (13b):

(14)



Here we see how both CP and TP follow the X′-schema (see (9)). We also see that TP is the complement of C and VP is the complement of T. The properties that characterize the C-T system have been extended to the nominal system, leading to the postulation of a D(eterminer) category that takes the NP as its complement (cf. Abney 1987, Horrocks & Stavrou 1987, Szabolsci 1983/1984, for some early approaches).

Having provided some evidence for the syntactic presence of categories such as C and T, mainly based on English, let us now turn to their morphological properties. Consider C, for example, which in English can be realized by means of free morphemes, such as *that, if, whether*. At the same time, question formation in (10a) and Germanic V2 in (13) show another kind of morphological realization of C, namely by means of moving a verbal element to the C position, giving rise to inversion. In other languages, complementizers are realized by

means of affixes, as is the case in Korean (a rigidly head-final agglutinating language – see Cinque 1999:53–54 and the references given there):

(15)

cap-hi-	si-	ess-	ess-	keyss-	sup-	ti-	kka
V-	Passive-Agr	Ant	Past	Epistemic	Agr	Evid	Q

'Did you feel that (unspecified argument) had been caught?'

Here (interrogative) C is realized by the suffix *-kka*. It is a matter for debate whether languages like Korean are C-final, or whether TP moves to the Specifier of CP (see Kayne 1994 on the latter possibility). It is also possible to find languages which have no realization (alternatively use a zero morpheme) for C; this is in fact an option in English under certain conditions (cf. *I said (that) John left*). The same pattern can be found in the T domain: elements indicating tense, mood, etc., can be realized as free morphemes, like English modals; as bound ones, like the passive, epistemic and evidential morphemes in (15); or receive no realisation at all, as in the English simple present. Similarly, D can be free (as in English), bound (as in Rumanian) or zero (as in Latin). Thus, to summarize the discussion so far, we see that functional categories are subject to cross-linguistic variation in their realization, are like normal lexical categories in that they can project their properties, but differ from normal lexical categories in being closed-class and (as is clear when we compare auxiliaries and main verbs in English) in being inherently 'defective' in various ways.

In the recent theoretical literature, it is possible to identify two main views on functional categories: one is to deny their existence, the other to accept them. We have already provided evidence against the first view, as grammatical information is relevant for syntax and appears to have configurational instantiations. Further evidence comes from the areas of typology, diachrony, and language acquisition. In particular, typological studies have shown that languages undeniably differ in word order and morphology. As we mentioned in the Introduction, positing functional categories, and making them the locus of cross-linguistic variation allows us to reduce these two axes of variation to one. Regarding diachrony, it is one of the aims of this book to provide evidence from grammaticalization for the syntactic presence of functional elements (see also the first section of this chapter). Finally, recent work on language acquisition, starting with Hyams (1986) and Radford (1990), has shown the relevance of functional categories in the development of grammars, as early grammars differ from the adult ones in the way functional categories are realized.

Having argued then that functional categories must be syntactically present in some way, various options are open for how this idea may be implemented – particularly in accounting for cross-linguistic variation in word order and

inflection. There are in fact various ways ranging from accepting a very small to a quite large number of functional heads. The first of these approaches could be called the 'What you see is what you get' (WYSIWYG) analysis. As its name implies, the only functional categories postulated as present in a given language, or even a given sentence, are the ones for which we see some kind of realization. On this view, it is a matter of parametric variation as to which functional heads are present in which languages. For example, Grimshaw (1997) proposes that structures can 'stretch', or be compressed, even in the same language, depending on the number of lexical items available. To illustrate this point, consider (16):

(16) I think [$_{VP}$ it rained] (Grimshaw 1997:410)

According to Grimshaw, the embedded clause here is just a VP, as no auxiliaries or complementizers are present to instantiate T or C. However, as the examples in (17a–c) show, this can't be right: if the subject was inside the VP, then substitution, fronting or deletion of the VP should also include the subject, contrary to fact:

(17) a. *do so*: I thought it rained and *(it) did so.
b. *VP Fronting*: I thought it rained and rain it did/*it rain did.
c. *VP Deletion*: I thought it rained and *(it) did.

Thus at least TP must be present.

Grimshaw's (1997) analysis has to account for cross-linguistic variation in a different way by imposing a different ranking of constraints across languages. The result of this is the reduction of functional heads at the expense of a proliferation of constraints. However, it is a conceptual consequence of the distinction between syntax and phonology that certain elements may be present at one level and absent at the other. We therefore expect that syntactic categories can be silent. So there is no conceptual advantage in a position like Grimshaw's. The next question then is whether this kind of view has empirical advantages. The answer seems to be negative for a number of reasons. Clearly, Grimshaw's analysis of (16) cannot be right, as we have just seen. More generally, though, WYSIWYG approaches complicate the statement of cross-linguistic variation, as we have to assume that variation lies in differing selections from a universally given pool of categories such that simple sentences may have quite different structures in different languages; but if functional categories have semantic content, then we might expect that simple declarative sentences all have the same category across languages – this assumption is natural from the perspective of

the mapping from syntax to semantics, and simplifies the task of the language acquirer. However, this is explicitly denied by the WYSIWYG approach, which may say, for example, that German main clauses are CPs while English ones are VPs. Also, grammaticalization is harder to understand on the WYSIWYG view. If grammaticalization involves the development of new functional material, it must be analysed as a structural change rather than a simple category change. Given the assumptions about learning and change articulated in the previous section, structural change is hard to account for (in fact, the innovation of whole phrases is impossible on this view – surely a desirable result in the light of the logical problem of language change). Categorial reanalysis is a natural change, though, as we sketched there (see the discussion of (6)).

Another possibility is to assume that functional categories are always present, but in a very restricted fashion. For example, Chomsky (1995, 2000) argues that categories like C, T and D are present as they carry clause-typing, temporal and referential information respectively. Other functional categories that were postulated in earlier versions of the theory, such as subject agreement (AgrS) and object agreement (AgrO) should be dispensed with, given that they are not conceptually necessary (see Chomsky 1995, 4.10 and section 4.5 of the present work for discussion). This looks like a viable option, but the problem is that this kind of reduced structure brings along other complications. In particular, in order to accommodate lexical material structurally we need to assume that there can be multiple Specifiers, subject to parametric variation. Although it is desirable to keep the number of functional heads to a minimum, this kind of approach has the consequence of becoming less restrictive.

There are at least two further possible approaches. One is to say that we accept a relatively large number of functional heads, provided we find empirical support for their existence. This is the line of reasoning followed by many recent studies in different ways. For example, Kayne (1994, 1998) accounts for certain ambiguities (e.g. in *I will force you to marry no one*), which in earlier frameworks are assumed to involve covert movement in order to determine scope (of the quantifier *no one* in the example just given), by means of overt movement to a number of functional projections. The existence of functional positions is justified on this basis, but there is no further attempt to justify their presence conceptually by means of identifying the properties that trigger movement in the first place.

Cinque (1999) argues for a number of functional heads based on the distribution of adverbs. Each position carries the property identified with the interpretation of the adverb, resulting in the following set (we give only the labels of

the categories; from left to right, each takes the maximal projection of the next as its immediate structural complement in the sense of the X′-schema in (9)):

(18)

$Mood_{Speech\ Act}$	$Mood_{Evaluative}$	$Mood_{Evidential}$	$Mod_{Epistemic}$	T(Past)T(Future)
$Mood_{Irrealis}$	$Mod_{Necessity}$	$Mod_{Possibility}$	$Asp_{Habitual}$	$Asp_{Repetitive(I)}$
$Asp_{Frequentative(I)}$	$Asp_{Continuative}$	$Mod_{Volitional}$	$Asp_{Celerative(I)}$	T(Anterior)
$Asp_{Terminative}$	$Asp_{Generic/progressive}$	$Asp_{Perfect(?)}$	$Asp_{Retrospective}$	$Asp_{Proximative}$
$Asp_{Durative}$	Voice	$Asp_{Prospective}$	$Asp_{SgCompletive(I)}$	
$Asp_{PlCompletive}$	$Asp_{Celerative(II)}$	$Asp_{SgCompletive(II)}$	$Asp_{Repetitive(II)}$	
$Asp_{Frequentative(II)}$	$Asp_{SgCompletive(II)}$			

This 32-head structure, as Cinque stresses, is a conservative estimate of the number of functional heads in 'TP'. No account is taken here of Negation Phrases or Agreement Phrases, for example. A similar approach is also taken by Manzini and Savoia (forthcoming) and Poletto (2000) who postulate a number of functional heads in the C and T domain, based primarily on the possible clitic strings found in Italian dialects and their interaction with verb movement, negation, particles, etc. Giorgi and Pianesi (1997) adopt another variant. They assume a universal set of functional features, all of which are in principle able to project – in this respect their approach is like those just mentioned. However, they also assume that – while there is a universal hierarchy of functional projections – features can 'scatter' over a structure in different ways in different languages. To put it another way, features can syncretize onto heads as long as the universal ordering (which is thus not a total ordering, in the technical sense) is not violated. Cinque (1999:133) criticizes this approach on the grounds that it is excessively complex (a special convention is needed to interpret syncretic heads); we will return to this issue in our discussion of markedness in Chapter 5.

The results of these approaches are no doubt enlightening and can complement an alternative view which attempts to identify functional heads on the basis of their interpretation. This is the view that we will pursue in the present book. In particular, we will argue that only those functional heads that have logico-semantic content can be present. This allows us to postulate a rather large number of functional heads, but at the same time the requirement for interpretability constrains what can be a functional head. For example, functional projections that play the sole role of being place-holders cannot exist. We sketch this approach in the next section, and return to it in detail in Chapter 5.

To summarize, in the present section we considered the reasoning behind accepting functional (grammatical) elements as syntactic entities. It is clear that

the presence of functional elements in syntax has considerable implications for typological studies, for the study of diachrony as well as acquisition. The next question of course is to identify what does and does not count as a syntactic functional head. We fully address this question in Chapter 5. In the next section, we will present in detail how we see functional heads providing the means to express cross-linguistic variation.

1.3 The nature of parameters: interface interpretation of functional categories

So far we have established a couple of main points. First, cross-linguistic variation is associated with functional elements and is restricted to the lexicon. Second, language acquisition is the process of setting parameters, while syntactic change is the result of changing (resetting) parametric values, in the sense discussed in 1.1; the parameter-setting device may, under certain conditions, fix a parameter p differently from the value assigned to p in the grammar that underlies the trigger experience. Finally, functional categories are syntactically present: they project their categorial features following the X′-schema in (9). On this basis, they are manipulated by syntactic operations such as Merge and Agree. What we need to do next is to clarify the nature of parameters, so that we can provide an account of grammaticalization. Since parameters are associated with functional heads, we need to specify the lines along which these heads may vary.

The approach we will outline here is based on Roberts and Roussou (1999), who aim at giving expression to the idea that movement, cross-linguistic variation and at least some morphophonological properties are reflexes of a single property of the computational system of human language (C_{HL}). This property of C_{HL} is driven by the interfaces, and is referred to as **interface interpretability**. The analysis takes the standard view of the interfaces as PF and LF, that is, the interfaces with the Articulatory-Perceptual and the Conceptual-Intentional systems respectively. Interpretability is the property of mapping a syntactic feature onto a PF or LF expression. To take a very simple example, the noun *table* maps onto a PF representation (/teibl/) and an LF representation, that is, its denotation ([[table]]). We cannot go into detail here as to the nature of the PF or LF representations, but it suffices to state quite simply that, in principle, any syntactic symbol may or may not be mapped onto a PF or LF representation. The lexicon provides the information determining the mapping. For ease of exposition at this point we could designate a syntactic symbol which has a PF mapping as +p, and a syntactic symbol which has an LF mapping +l. So *table*

is both PF- and LF-interpretable. In fact, we can observe that the lexical entries of lexical categories, such as Nouns and Verbs, always contain a specification +p, +l.

Consider next a functional element such as C, which, as we mentioned in the previous section, provides information about clause-typing, among other things. This kind of information contributes to the interpretation of the sentence, so we take C (or more precisely the features associated with C) to be LF-interpretable, that is +l. We saw in the previous section that the realization of C is subject to parametric variation: a matrix declarative C in German is realized by means of a verbal element, which is partly responsible for the V2 construction, while its English counterpart receives no such realization. Instead the matrix declarative C in English is not spelled out, or alternatively is spelled out as zero. Another example of parametric variation was discussed in section 1.1 in relation to the 'pro-drop' parameter. Let us assume that AgrS is the position associated with the nominal features of the subject. As such it receives an interpretation at LF (+l)[1]. Its realization, however, differs across languages. For example, English AgrS requires an overt subject, while its Italian counterpart allows for a null subject. We see then that functional elements are not necessarily +p. Variation in ±p leads to cross-linguistic variation in which functional categories are overtly realized, as we will see in detail below. In general, then, we see that functional categories may be defined as that class of syntactic categories which is not obligatorily +p.

Among the functional features, Q, WH, Neg, T and D, at least, are LF-interpretable, that is, +l. These features clearly contribute to the interpretation of any phrase-marker they appear in. We assume that UG contains a vocabulary of substantive universals, which are realised as functional features in every language. These are the interpretable features. On the other hand, the ±p property varies across languages. In fact, *pace* Chomsky (1995, 2000, 2001), we do not postulate uninterpretable functional categories or features. We believe that it is possible to maintain that all such features are LF-interpretable. If so, then there are no [−l] features.

The ±p and ±l properties are lexically determined, and as such are listed in the lexical entries of morphemes (cf. Cormack & Smith 1999 for a slightly similar approach). Assuming that lexical items are bundles of features, we can say that any category with N and V features is always +p, +l, while functional

1. Chomsky (1995, 2000) argues against the postulation of an AgrS category, on the grounds that phi-features are interpretable for nominals and not for verbs. However, if we take AgrS to correspond to a position that encodes the nominal features of the subject in the clause structure, then its presence becomes legitimate as these features are clearly interpretable. See section 4.5.

categories are +l, ±p. It is interesting to observe that functional heads, although they may have interface content, always have relatively impoverished content as compared to lexical heads. For example, a good approximation of the content of 'verbal' functional material (auxiliaries, aspect, tense) is that it is lacking in argument structure. Similarly, +p functional heads are almost always phonologically 'light', often lacking in stress, or failing to meet the criteria for minimal wordhood (see McCarthy & Prince 1986), as is the case for monomoraic *the* and *a* in English. It may be then, that functional elements fall below certain threshold values for phonological and semantic content even when they have interface properties. This idea may also contribute to an account of why grammatical systems vary and change, since the crucial PF information is presented in a 'weak' form. We return to this point in Chapter 5.

Let us notate a functional feature F that requires a PF realization as F*. Parameterization is seen as the random assignment of the diacritic * to features typically associated with functional heads. Where the diacritic is assigned to a feature, that feature, F*, must have a PF realization. Again, * is assigned to F in the lexicon, following Borer's (1984) idea that parametric variation is a facet of the lexicon. The overall conception of the lexicon, then, is that it contains the following elements:

(19)
a. Lexical items, specified as ±V, ±N, with PF and LF properties given
b. Substantive universals encoded as interpretable features of functional heads
c. *assigned in a language-particular fashion to (b)

The only variant property is the assignment of *. The diacritic * does not apply to lexical heads, as these seem to be inherently associated with phonological features. The diacritic * is the expression of a relation between functional features and morphophonological matrices (overt or zero morphemes). Notice that under this view of variation there is no selection among the universal set of features. In other words, all languages have the same set of functional features; what varies is whether and how these features are realized in PF. This seems to be the null hypothesis and is in principle open to falsification, although Cinque's (1999) results suggest that the null hypothesis is correct. Thus there is no parametric variation in this respect (see also our discussion of what we called WYSIWYG models in the previous section).

Let us now turn to the PF realization of F*. This can be achieved in two ways: by Move or by Merge. Which option is taken depends on what the lexicon makes available, but the most economical is always preferred. For this reason, Merge is always preferred over Move. If the lexicon provides a

morphophonological matrix for F*, then this matrix will be F*'s realization, and Move is unavailable. Conversely, if the lexicon has no such matrix for F*, material from elsewhere must be moved to F (subject to the usual constraints on movement). Alternatively, we can view * as the morphophonological matrix for F; in this way, its cross-linguistic arbitrariness becomes completely natural – as de Saussure ([1916] 1972) pointed out, the relationship between linguistic elements and their morphophonological instantiation is arbitrary. We thus see a further dimension of parametric variation along the Move versus Merge axis. Since these are the only ways of associating lexical material with syntactic positions, they represent natural options.[2] So we have the following system of parametric variation:

(20) a. F* ? Yes/No
b. If F*, is it satisfied by Move or Merge?

The least economical option is Move. Following Clark and Roberts (1993) we might therefore expect this to represent the marked option, the one for which robust evidence must be available to language acquirers. At the same time, Merge would be the less marked option. This is crucial for the account of grammaticalization we will offer in the following chapters.

Before concluding this section, let us briefly illustrate how this system of parametric variation works (for further details see Roberts & Roussou 1999, 2002, Roberts 2001). We will give two examples: yes/no questions and wh-questions. Consider first yes/no questions: the feature responsible for giving a clause this interpretation is identified as Q. Since this is a clausal property, it is natural to associate Q with the head position of the clause, that is, C. We then observe the following variation:

(21) a. Did John see Mary? (English: Q^*_{Move})
b. A welodd John Mary? (Welsh: Q^*_{Merge})
c. Jean a vu Marie? (Colloquial French: Q is silent)

The PF realization of Q varies as a function of what the lexicon makes available: English has Q*, but no Q particle, and so movement (of T) is chosen. Welsh has Q* and a Q particle (*a* in (21b)), and so movement is blocked by the more economical Merge. In Colloquial French (21c), Q has no PF realization. In this case, interrogative force has no overt syntactic realization and is marked purely by intonation. In other words, taking F to be Q, we find the two possible realizations of Q* by Move and Merge in English and Welsh respectively. The Q option is attested in Colloquial French.

2. There is one further possible option, namely that F* may be associated with a morphophonological matrix which is a syntactic affix, and which hence triggers both Move and Merge, following the Stray Affix Filter (or whatever constraint this follows from). From an economy perspective, this option is equivalent to Move (on the assumption that Merge is costless, cf. Chomsky 1995), but see Chapter 5.

Consider next wh-questions. We know that in some languages, wh-phrases are fronted to clause-initial position, as in English, while in other languages, they remain *in situ*, as in Chinese (cf. Cheng 1991 for a typological discussion). The cross-linguistic pattern shows more variation than that, but for present purposes we restrict our attention to a few simple cases:

(22) a. Who did John see – ?
b. Hufei chi-le *sheme* (ne) (Chinese, Cheng 1991:112)
Hufei eat-asp what Q_{wh}
'What did Hufei eat?'
c. Mona shaafat *meno*? (Iraqi Arabic, Wahba 1991:253)
Mona saw whom
'Who did Mona see?'

As (22a) shows, English has wh-fronting, while Chinese (22b) and Iraqi Arabic (22c) illustrate wh-in-situ. Moreover, in main-clause questions English requires subject–auxiliary inversion in addition to wh-fronting:

(23) a. *Did John see who?
b. *Who John saw?

Regarding yes/no questions we suggested that the auxiliary *did* realizes the Q feature of C. As the ungrammatical (23a) shows, the wh-phrase must also front. Let us assume then that while *did* realises the Q feature, it does not identify the clause as a wh-question. In other words, the wh-C must also be spelled out as such. Given that the wh-feature is part of a DP, the whole wh-phrase is fronted and realises wh-C*. Similarly, absence of auxiliary inversion in (23b) gives rise to ungrammaticality as the Q* is not spelled out. We also find wh* in constructions which are not interrogative, and which as such lack Q. This is the case for exclamatives, for example. Here, subject–auxiliary inversion is impossible, and wh-movement obligatory, as (24) illustrates:

(24) a. What a nice guy he is!
b. *What a nice guy is he!
c. *He is what a nice guy!

This pattern follows straightforwardly in our system from the fact that these are non-interrogative wh-constructions, and so Q* is absent but wh* is present.

Turning now to the Chinese example in (22b), we notice that there is no wh-movement, as the element *sheme* remains *in situ*. That might lead us to the conclusion that wh-C in Chinese does not require an overt realization. However, this is wrong for a number of reasons. First, Chinese, unlike English, does not have dedicated wh-words; the element *sheme* in (22b) is actually an indefinite which is also found in a number of contexts, such as yes/no questions, negated clauses, etc. (cf. Aoun & Li 1993). The interpretation of *sheme* depends on the interpretation of the clause. Second, the question particle *ne* in (22b) is in complementary distribution with the yes/no particle *ma*:

(25) Qiaofong mai-le *sheme* ma
Qiaofong buy-asp what $Q_{y/n}$
'Did Qiaofong buy anything?'

As (25) shows, when *ma* is present the sentence is interpreted as a yes/no question and *sheme* has the reading of *anything* and not of a wh-element. Given the complementary distribution of these two particles, we can argue that wh-C in Chinese is also * (bearing also in mind that Chinese is C-final). However, it differs from English in that wh-C* is realized by means of Merge, that is, inserting the dedicated wh-particle *ne*. Since the wh-feature is spelled out by Merge, Move is pre-empted. This analysis derives Cheng's (1991:30) generalization that languages with dedicated wh-particles lack wh-XPs and wh-movement.

The final example in (22c) comes from Iraqi Arabic, which, like Chinese, does not show wh-movement, but, unlike Chinese (and like English), has a distinct class of wh-words. That is, the word *meno* can only be used as a wh-element. According to Ouhalla (1996) wh-words in Iraqi Arabic consist of the wh-part plus a pronominal clitic: *men-o* = 'who+him'. In that respect they crucially differ from their Chinese counterparts, which are simply indefinites bound by any compatible operator. Also, the Iraqi Arabic sentence does not make use of a dedicated wh-particle. This is because the wh-feature is morphologically realized on the relevant DP, once again in accordance with Cheng's (1991) generalization. On the basis of this pattern, we could argue that wh-C in Iraqi Arabic does not receive an overt realization.[3] We thus see that the three options of wh-C*Merge, wh-C*Move and wh-C are instantiated by Chinese, English and Iraqi Arabic respectively. This kind of approach has the clear advantage of allowing us to capture a wider range of cross-linguistic variation in a more principled way than the standard formulation of the wh-parameter in terms of the presence versus absence of movement.

3. Iraqi Arabic shows optional fronting, as in (i) (Wahba 1991, Ouhalla 1996):

(i) *meno* Mona shaafat?
who Mona saw
'Who did Mona see?'

We will assume, following Ouhalla (1996) and Cheng (1991), that optional wh-fronting of this type is an instance of a reduced cleft construction. Thus the structure in (i) is not a counter-example to the analysis suggested in the text.

The above constructions are just an instance of how the notion of interface interpretability can be used to account for parametric variation. The crucial point is that features must receive an interface interpretation. Thus we predict that there can be no features that receive no interpretation at either interface. That is, we exclude the presence of -p, -l features. This class of features would correspond to Chomsky's (1995, 2000, 2001) covert uninterpretable features. Assuming that all features receive an interpretation, we are able to locate parametric variation in the lexicon and in particular to restrict it to the ±p property of functional heads. A further important implication of this system is that it predicts that all instances of movement are overt, that is, there is no covert movement (cf. also Brody 1995, 1997, Kayne 1998, among others). This has important implications for the structure of grammar, which we will not investigate further here (but see Chomsky 2000 for a recent discussion). At the same time this approach brings PF and LF closer, as the former can be used as an indication of the kinds of relations that are established at LF.

1.4 Conclusion

In the present chapter we have introduced the main ideas that we will assume. In particular, in section 1 we introduced the logical problem of language change, formulated within the Principles and Parameters framework. As we pointed out, language acquisition is the process of setting parameters, whereas language change is the result of changes in the parametric settings. This view of language change links it closely with language acquisition. As Lightfoot (1979) pointed out, changes arise when a population of learners converges on a grammar which is distinct from the grammar that creates the input to learning (the trigger experience, or cues, in the terminology introduced in 1.1). In other words, the final state of acquisition may not result in full convergence with the adult grammar. We suggested that this happens when the trigger for a particular parameter value is obscure or ambiguous. In this case, parameter setting is facilitated by the learning device which has a built-in preference for simpler representations. This then opens the way to a solution to the logical problem of language change.

In section 2 we focussed on the nature of functional categories and argued that they have a syntactic status, as they have an effect on the syntactic relations and are visible to syntactic operations. We also argued against an approach that

parameterizes the number of available functional heads cross-linguistically (the 'WYSIWYG' theories) or restricts the number of available functional heads to a considerable extent (Chomsky's (1995, 2000, 2001) approach). The view we will pursue in the present book argues for the availability of a number of functional heads.

Finally, in the last section we outlined the theory of parameterization that we will assume, following Roberts and Roussou (1999). The main idea is that there are only interpretable features. Interpretability holds for both interfaces, LF and PF. Features associated with functional heads are LF-interpretable, but may not be PF-interpretable. It is the latter property that gives rise to cross-linguistic variation. Thus a feature F may be left unrealized (i.e. not spelled out) or be realized in one of the following two ways: by Merge (lexical insertion) or Move (attracting another morpheme). Which option is chosen depends on what the lexicon makes available. In this system parametric variation arises with respect to which features are spelled out and how. In the following chapters we will show that this approach has important implications for syntactic changes, and in particular for grammaticalization, which can be viewed as change from the Move to the Merge option.

Having outlined the basic assumptions of our approach, we next turn to a more detailed presentation of the relevant data. The next three chapters are devoted to applying the ideas presented here to a range of case-studies in grammaticalization. We take each main functional domain in turn: the T system, the C system and the D system. Our focus in these chapters is empirical; in Chapter 5 we will return to a more detailed and empirically informed discussion of some of the issues raised here.

2 *T elements*

2.0 Introduction

The purpose of this chapter is to provide empirical evidence for the claim that grammaticalization involves reanalysis of functional categories. The central idea is that whenever grammaticalization takes place, the content of at least one functional category is reanalysed, in such a way that new morphophonological realizations of functional features are created. In the notation of Chapter 1, section 1.3, new cases of F^* develop for some feature F (usually, but not always, F^*_{Merge} is innovated – see below and Chapters 3 and 4 for empirical evidence). This can mean that a lexical item or class of lexical items is reanalysed as functional, or that one functional category develops into another. We will see cases of both kinds in what follows. Crucially, our approach to grammaticalization implies that it is not a structural change: functional structure is present both before and after grammaticalization takes place; what changes is the way the features associated with functional heads are realized. More precisely, assuming a universal hierarchy of functional heads, as mentioned in Chapter 1, section 1.2 (see also Cinque 1999), the change involves the overt realization of these heads.

The chapter is organised as follows: section 2.1 deals with the grammaticalization of V elements to T markers, focussing on the development of modals in the history of English. This has been treated as a typical case of grammaticalization whereby a lexical verb is reanalysed to an auxiliary element. Section 2.2 discusses the development of the Romance futures from the infinitive + *habeo* construction, as a case where a lexical verb becomes an auxiliary and is finally reanalysed to a verbal affix. Finally, section 2.3 deals with the development of the future particle *tha* in Greek from the volitional verb *thelo*, or more precisely the sequence *thelo* + *na V*. Although all these cases are very well-trodden ground, we show that our treatment of them provides support for our approach. To this end, we show that the principal generalization is that categorial reanalysis always involves reanalysis of movement. We then consider a

theoretical generalization that follows from the empirical one: is it possible that grammaticalization, as reanalysis of movement, is isomorphic to movement? This idea would impose a strong restriction on what grammaticalization can be. By looking at the relations among functional categories, we observe two things: (i) that the diachronic movement of a given morpheme, possibly tracked over many centuries through successive reanalyses, is always 'upwards' in the structural hierarchy of functional categories (here Cinque's 1999 work becomes particularly relevant); (ii) that much of the allegedly continuous or cline-like nature of grammaticalization is due to multiple 'lexical splits'; as we will show, the different readings attributed to a single lexical item correspond to different positions in which it may be merged in the clause structure (see Poletto & Benincà 1997 and Simpson and Wu 2001 on this). We will also suggest a way in which it is possible, using the universal base proposed in Kayne (1994), to predict that new affixal material can only arise from head-movement constructions. Throughout this chapter and the following two, we largely leave aside questions of semantics and phonology – these will be dealt with to a greater or lesser extent in Chapter 5. Our goal here (and in Chapters 3 and 4) is to establish a clear set of empirically well-founded generalizations regarding the syntax of grammaticalization. Once this is clear, we can tackle the wider issues in Chapter 5, beginning with the explanation of grammaticalization as a syntactic change in terms of the assumptions put forward in Chapter 1, section 1.1.

2.1 *From verb to auxiliary: the development of English modals*

The well-known development of the English modal auxiliaries is a fairly clear case of grammaticalization in which what were once fully verbal elements underwent a category change and became auxiliaries. The basic evidence that modals are syntactically distinct from main verbs in Modern English (NE) is very well known, and we recapitulate it in (1):

(1) a. Modals lack non-finite forms:
*To can swim is useful.
b. Modals cannot be iterated:[1]
*He shall must do it.
c. Modals lack complements of all types (except bare infinitives):
*I shall you a penny.
d. Modals are in complementary distribution with *do*-support and always precede *not*:
*I don't can speak Chinese.
*Do you can speak Chinese?
*I not can speak Chinese.
e. Modals always move to C in inversion contexts (cf. (14) of Ch. 1):
*How many languages (do) you can speak?
f. Modals, unlike main verbs, can license VP fronting (and also VP ellipsis, as in (10c) of Ch. 1):
Win the election, I thought she would (*win) —.
g. Modals, unlike main verbs, can phonologically contract:
We can fish. — ambiguous ('we are able to fish' or 'we put fish in cans')
We c'n (/kən/) fish. — unambiguous (only 'we are able to fish').

In all these respects, modals are distinct from main verbs, which have non-finite forms, allow iteration (via clausal complementation), have a variety of complements (as a matter of selection/subcategorization), cannot precede clausal *not* or invert (but instead require *do*-support in the relevant contexts if no other auxiliary is present), cannot license VP fronting and cannot contract. It is worth clarifying at this stage that not all verbs with a modal reading exhibit the properties in (1). The verb *ought* is a clear example, as despite its modal character, it syntactically behaves like a lexical verb in some varieties (e.g. *I ought to go, I didn't ought to go*).

The properties of NE modals can be accounted for by merging them in the verbal functional system, in a position inaccessible to lexical verbs in NE (but accessible to other auxiliaries *do, have* and *be*), since main verbs do not raise into the functional system in NE (see Roberts 1985, Pollock 1989, and below on this point). For the moment, let us call this position T, and assume that T is the only functional category between C and VP. If we say that modals are members of T, then we may expect them to be sensitive to particular properties involving T (or the C-T relation), such as finiteness – hence the finiteness restriction is naturally stated. The lack of iteration could be accounted for in terms of the uniqueness of T: since there is just one T present, there can only be one modal in each clause. However, the availability of multiple-modal data like that mentioned in note 1 suggests that the uniqueness of T is not sufficient

1. Except in Scots, certain dialects of Northern England, and Southern US dialects, where double and perhaps triple modal constructions can be found (Brown 1992, Roberts 1993a:333, n. 3, and references given there). Cinque (1999:54, 78f.) suggests that sequences of modals fit into his clausal hierarchy. On the other hand, Battistella's (1991) evidence that the first modal in a sequence such as *He might could do it* is 'spurious' and may be an adverb is problematic. Roberts (1993a:317), citing Plank (1984) (who in turn cites Ščur 1968), points out examples from Leicestershire English and Scots English where root modals retain non-finite forms. This may be the same fact as that discussed by Brown, to the extent that, where the first modal is a true modal, the second modal must be non-finite for a sequence to be possible.

to predict non-iteration. Instead it should be handled in the same way as the lack of non-finite forms: since modals always require a following bare infinitive and themselves have no infinitive forms, they cannot iterate. This is a desirable result as it allows us to have more than one position for modals and predict the availability of multiple-modal sequences for those modals only that have an infinitival form. We will come back to this point shortly.

At earlier stages of the language, prior to approximately the sixteenth century, none of the above properties characterized modals as a class distinct from lexical verbs, although Warner (1993:111f.) suggests that even in Old English (OE) the ancestors of some modals and *be* may have been able to license VP ellipsis. The following examples illustrate the lack of the relevant distinctions in Middle English (ME). Examples (2a–c) illustrate this point with modals, and (3a–b) with main verbs; this is the best way to indicate the properties that have been lost as the two classes have become distinct (we leave VP fronting and auxiliary reduction aside – see Plank 1984 on the former and Warner 1993:207–208 on the latter):

(2) a. *Non-finite modal*:
but it sufficeth too hem **to kunne** her *Pater Noster*, ...
but it suffices to them to know their *Pater Noster*, ...
(?c1425 (?c1400) *Loll. Serm.* 2.325; Denison 1993:310)

b. *Iteration of modals*:
Who this booke **shall wylle** lerne ...
He-who this book shall wish learn ...
'He who wishes to master this book.'
(c1483 (?a1480) Caxton, *Dialogues* 3.37; Denison *ibid.*)

c. *Complementation*:
euerych bakere of þe town ... **shal to the þe clerke of þe town a penny**
every baker of the town ... owes to the clerk of the town a penny
(a1400: Usages of Winchester (Engeroff), p. 64; Visser 1963–1973, §549; Roberts 1993a:313)

(3) a. *Main verbs preceding 'not'*:
if I **gave not** this accompt to you
if I gave not (=didn't give) this account to you
(1557: J. Cheke, Letter to Hoby; Görlach 1991:223, Roberts 1999:290)

b. *Main verb inversion*:
How **cam'st** thou hither?
How camest thou (did you come) here?
(1594: Shakespeare, *Richard III*; Roberts *ibid.*)

Both Lightfoot (1979) and Warner (1993:100–102) argue that the ancestors of the modern modals, the 'pre-modals' in Lightfoot's terms, were main verbs in OE and ME, although most of them were members of a particular morphological

class, the preterit-present verbs. This class, which in addition to the pre-modals included a small number of now-obsolete items (see Warner (1993:140–144) for detailed discussion and illustration), has present forms which derive historically from an Indo-European perfect form. As Lightfoot (1979) pointed out, the consequences of this are (a) that there was never a distinct third singular ending *-(e)th* or *-(e)s* (for example, the 1sg and 3sg OE forms of 'shall' are *sceal*), and (b) that the past tense was highly irregular. Lightfoot and Warner both suggest that these morphological peculiarities played a role in singling out the pre-modals as a subclass.

Whatever other complements they may have had at earlier stages, modals were able to take infinitival complements at all times, although in this respect too they may have been unusual, since, with the exception of *agan/ought*, they are rarely found with *to*-infinitives – see Warner (1993:136–139) for discussion and mention of one or two apparent ME exceptions. We take infinitives to be at least TPs: since finiteness is a property of T, this element must be present in order to define a clause as non-finite (we assume that the presence of VP is uncontroversial). Now, as shown by Roberts (1985, 1993a:246–255), earlier English (until at least the sixteenth century) had productive V-to-T raising in finite clauses. This is clearly shown by examples like (3d) where V precedes clausal negation, and (4) where the verb and its nominal direct object are separated by an adverb or a floating quantifier (see Pollock 1989 for discussion and justification of these tests):

(4) a. The Turkes ... **made anone redy** a grete ordonnaunce.
The Turks ... made soon (=soon prepared) a great ordnance.
(c1482: Kaye, *The Delectable Newsse of the Glorious Victorye of the Rhodyans agaynest the Turkes*; Gray 1985:23; Roberts 1993a:253)

b. In doleful wise they ended both their days.
(1589: Marlowe *The Jew of Malta* III, iii, 21; Roberts *ibid.*)

If infinitival complements contain T and main clauses feature V-to-T movement, then simple pre-sixteenth-century examples containing a pre-modal and a bare-infinitive complement like (5) must have had a biclausal structure like (6):[2]

2. The structure in (6) glosses over a number of complexities that are not directly relevant here. In particular, we treat the adverb *sone* as adjoined to TP. In fact, it is much more likely to be in a topic position in the CP system (Cinque (1999:96f.) argues that adverbs like *soon* occupy the Specifier of an AspectP which is rather low in the clause structure, certainly lower than the position occupied by epistemic modals like 'may' in (5a), hence the adverb must have been topicalized in this example). The structure in (6) does not include any of the elaborated functional structure proposed by Cinque, collapsing it all as TP in both clauses. Moreover, we gloss over the question of the nature and presence of a CP layer in the lower clause, the VP-internal subject hypothesis and the nature of the empty category in the subject position of the lower clause. None of these points affects the present discussion.

(5) a. Sone hit mæi ilimpen
soon it may happen
(a1225 (?a1200): Lay. *Brut* 2250; Denison 1993:299)
b. þou mai haue childer
You may have children
(a1425 (?a1350): *7 Sages* (2) 2843; Denison 1993:300)

(6) $[_{TP}$ Sone $[_{TP}$ hit mæi $[_{VP}$ $t_{mæi}$ $[_{TP}$ T $[_{VP}$ ilimpen]]]]]

Roberts (1993a:313–314) provides evidence that pre-modals in ME allowed both raising and control infinitives, that is, that the empty lower subject in a structure like (5) could be either a DP trace (raising) or PRO (control). (6) arguably illustrates raising, as expletives like *(h)it* are unable to participate in control relations. The clearest control examples involve dative experiencers in the main clause, since raising to an indirect-object position is impossible:

(7) a. Mee moste nedys been dampned for this
Me must needs (=I must) be damned for this
(1455: *Speculum Misericordie*, 251, Visser 1963–1973, §1715; Roberts 1985:38)
b. hwi mi ouh and hwi me scal iesu crist luuien
why me ought and why me shall J. C. love
(*Ancr. R.* (EETS 1952) 6, 23, Visser 1963–1973, §1712; Roberts 1993a:314)

There is some reason to think that pre-modals were restructuring verbs, in the sense of Rizzi's (1982) analysis of a class of Italian modal and aspectual verbs, that is, verbs which obligatorily or optionally triggered clause-union with their infinitival complement (see also Aissen & Perlmutter 1983). First, it might be that the upstairs case of the experiencer was determined by the downstairs verb: Warner (1993:122f.), following Denison (1990), discusses some possible instances, although his evidence seems rather equivocal. Second, van Kemenade (1993) gives evidence that OE pre-modals triggered verb (projection) raising, a process of reordering in the verbal cluster characteristic of verb-final West Germanic varieties which, since the earliest analyses (Evers 1975), has been related to restructuring (see also Rutten 1991). Third, it is known that verb raising triggers are incompatible with various particles in the complement (*om* in Dutch, both *um* and *zu* in German); in this light, the fact that pre-modals consistently select a bare, *to*-less infinitival may fall into place. Fourth, they exhibit long object-shift of negative objects (Beukema & van der Wurff 2000). We will not take a definite view on whether the pre-modals were restructuring verbs, although this seems to be likely. We will return below to one potentially important consequence of this idea.

The structure in (6) became a monoclausal structure like (8) when the modals were reanalysed:

(8) $[_{TP}$ Soon $[_{TP}$ it may $[_{VP}$ happen]]]

Roberts (1993a:310f.) dates the reanalysis from (6) to (8) as taking place early in the sixteenth century; Warner (1993:198f.) gives c1500 as the time of the loss of non-finite forms of pre-modals, an important consequence of this change, as we will see below. As argued in Chapter 1, this change is clearly favoured by the conservative nature of the learner as it involves the elimination of a movement operation: V-to-T movement of the modal. A consequence of the loss of this movement is that there is no longer evidence for a biclausal structure – only one instance of T is triggered after reanalysis (and this is now T^*_{Merge}, so the element merged in T provides it with a realization). In this way, the loss of movement leads to grammaticalization (modals are reanalysed from V to T), and the grammaticalization entails a reanalysis of the earlier biclausal structure in (6) as the monoclausal structure in (8). The root cause of this latter reanalysis is the loss of movement. So we do not need to ask what caused the change. What we need to see is what prevented the change from taking place sooner. In other words, what was present in the trigger experience of acquirers until c1500 that provided robust evidence (a robust cue, in the terminology of Lightfoot 1998) for treating the modals as verbs and/or for treating the structure of (5) as biclausal? We're looking for a cue for two Ts (biclausal structure) and/or for the verbal nature of the modals, a piece of evidence that was somehow lost or obscured around 1500.

Following Roberts (1993a), we propose that the causal factor was morphological. Indeed Lightfoot (1979) and Warner (1993) point out that the pre-modals as a subclass were 'opaque' as main verbs in a variety of respects, but the factors contributing to this opacity go back to OE and did not alter significantly around 1500 – see Warner (1993:198ff.) for discussion, and some of the points made below. By or shortly after 1500, the former infinitive ending *-e(n)* had disappeared (see Roberts 1993a:261). At the same time, *for NP to VP* constructions appear (Lightfoot 1979:186ff.). We can analyse this as a change in the parametric property of T. Earlier T attracted V with the relevant morphology (T^*_{Move}). Assuming standardly that *for NP to VP* constructions show that *to* must be in T (cf. also Pollock 1989), they are evidence for T^*_{Merge} in infinitival contexts (but see Chapter 3, section 3.3, for an alternative view, which does not affect the present discussion). Now, as long as infinitives occurred with the *-e(n)* ending, there was clear evidence for the lower T: this ending could not instantiate the higher T as modals were tensed, and so they instantiated the higher one.

Examples like the following were thus unambiguous evidence for two Ts, and therefore for a biclausal structure:

(9) nat can we se**en** . . .
Not can we see
'we cannot see'
(c1400: Hoccleve *The Letter of Cupid* 299, Gray 1985:49; Roberts 1993a:261)

So once the infinitival ending was lost, (6) was reanalysed as (8), thanks to the fact that there is no further evidence for the lower T, and hence no evidence for two Ts. This in turn means that there is no evidence for a biclausal structure, and hence for V-to-T movement. The crucial point to notice here is that while prior to the loss of the infinitival ending modals could be monoclausal (due to restructuring), after the loss of *-en* they had to be, that is, they became incompatible with a biclausal structure. This analysis has the advantage of accounting for the peculiar status of NE modals: NE is the only Germanic language with such a syntactically defined class, and it is the only Germanic language lacking an infinitival ending. The latter fact also correlates with the existence of the *for NP to VP* construction in NE but nowhere else in Germanic, and arguably with the non-existence of a particular kind of causative in NE (see Roberts 1993a:321).

Once the modals were grammaticalized as elements of T their NE properties emerge. In particular, they lose their argument structure,[3] and therefore the possibility of any form of structural complement other than a VP (which of course looks just like a bare infinitive), they lose their non-finite forms, and they take on complementary distribution with supporting *do* (which also underwent the same reanalysis; Denison 1985, Roberts 1993a:295, Warner 1993:201). It appears that these changes take place early in the sixteenth century (cf. Warner (1993:198ff.) for a thorough discussion). However, the picture is complicated somewhat by the existence of lexical splits whereby a given pre-modal divides into a grammaticalized element (member of T) and a full verb (member of V). For example, Warner (1993:202) describes the development of *can* as follows: 'In CAN, the sense "learn" becomes established in the spelling *con* as a distinct verb taking regular inflections.' We thus have can_T, a modal without argument structure and non-finite forms, and con_V a transitive verb meaning 'learn' with a full range of regular forms. In Standard English, *need* and *dare* survive to the

3. On the possibility of developing argument structure see Vincent's (1999) discussion on prepositions.

present as T-V doublets;[4] Ščur (1968) (cited in Plank 1984) points out dialectal examples where *can* and *will* survive as lexical verbs (see note 1).

Later in the sixteenth century (or possibly later still – see Tieken-Boon van Ostade 1987, Warner 1997:382–383, Lightfoot 1999:163), V-to-T movement of main verbs was lost (Roberts 1985, 1993a:246ff., 1999), which meant that only auxiliaries, that is, members of T, could precede clausal negation and move to C in interrogatives and other inversion contexts. This created the situation we observe in English today: a rigid separation exists between lexical verbs and modal auxiliaries. So we see how modals may have been reanalysed in the early sixteenth century from being verbs which, like all other finite verbs in the language at that time, moved to T to being merged in T. The crucial factor that led to this reanalysis was the loss of infinitival morphology, and the change in the modals took place at almost the same time as (or perhaps a generation later than) this.[5]

However, if we take into account a wider range of data, and the possibility of a richer functional structure, the picture becomes more complex and more intriguing. First, Warner (1983, 1993) gives evidence that at least some pre-modals may have started 'leaking' into the functional domain from much earlier than the sixteenth century. He shows that OE *mot, dearr* and *sceal* do not have attested non-finite forms (see his 1993, Table 6.3, p. 145). Warner goes on (p. 147) to formulate the following generalization for ME (his (3)):

(10) Preterit-present verbs subcategorized for the plain infinitive which denote necessity, obligation and related notions of futurity are finite only.

The generalization covers *mot, shal, þarf, mun* and *dar*, the core deontic modals of ME.

Taking our cue from Benincà and Poletto's (1997) important work on the Italian modal *bisogna* (be necessary), which we discuss in more detail below, we propose to account for Warner's generalization, and the general 'leakage' of pre-modals into the functional system which he documents, in terms of

4. With the added complication that modal *need* is a polarity item in present-day English:
i. *John need do that.
ii. John needn't do that.
Also, *need* must be interpreted inside the scope of negation in (ii).

5. Warner (1993:199) makes a useful comparison of the incidence of non-finite *can* and *may* in Caxton (1422–1491) and More (1478–1535), the latter making markedly less use of such forms than the former. It would be revealing for the account of the reanalysis of the modals offered in the text to check the incidence of infinitival *-e(n)* in Caxton.

Cinque's (1999) proposals for the functional structure of the clause. Cinque (1999:106) proposes a structure featuring some thirty functional categories, ten of which are mood or modal heads. The substructure which concerns us here is the following:[6]

(11) $\text{Mod}_{\text{Epistemic}}$ T(Past) T(Future) $\text{Mood}_{\text{Irrealis}}$ $\text{Mod}_{\text{Necessity}}$ $\text{Mod}_{\text{Possibility}}$ . . . Mod_{Root}

Suppose that the OE and ME modals in question were, in the relevant interpretations (i.e. 'necessity, obligation and related notions of futurity' as in (10)), able to be merged directly into the relevant functional heads. If we take the basic difference between lexical verbs and verbal functional heads to be the possession of argument structure, then we can think that merger directly into the functional system correlates with the absence of argument structure. A complication for this view stems from the fact that dynamic modals (or root modals) may be associated with (possibly defective) argument structure (see Jackendoff 1972, Zubizarreta 1982). If we want to follow Cinque's (1999) system we have to assume that 'necessity, obligation and related notions of futurity', in Warner's formulation, are notions that can be structurally expressed either through argument structure or by scopal properties of functional heads (we will elaborate on this below). The latter option is always preferred as it creates simpler structures (only one functional hierarchy rather than two). This is why the ME pre-modals that fall under (10) were able to be grammaticalized, in these interpretations, early. From the Mood/Mod position they moved higher, at least as far as the highest T, just as all finite verbs did at these periods (this assumption is unchanged from the discussion above). So this is grammaticalization of these modals on these interpretations, and is motivated exactly like the general grammaticalization of the modals described above: merging a modal higher in the structure economizes on movement steps, and so is preferred by the learner.

However, merging these modals directly rather 'high' in the functional structure meant that certain properties that had to be licensed (or checked) by lower functional heads could not be licensed. In particular, below the lowest modal head is a series of aspectual heads (cf. Cinque 1999, Chapter 2). It is plausible to suppose that participial morphology is licensed there. In this way, then, we can explain the absence of participial forms of the relevant modals; our explanation exactly parallels that offered by Benincà and Poletto (1997) for the morphological defectivity of Italian *bisogna*. We must also assume, following Benincà and Poletto, that infinitival morphology is checked lower in the structure than the modal heads. We will come back to the structure in (11) and modify it accordingly shortly. For the time being it suffices to show that the different interpretations of modals can be taken to correspond to different (functional) heads in the clause structure.

6. Cinque actually labels Mod_{Root} as $\text{Mod}_{\text{Volition}}$, but elsewhere (p. 90), he provides evidence for 'the postulation of three distinct root modal projections, in the order: $\text{Mod}_{\text{Volition}}$ > $\text{Mod}_{\text{Obligation}}$ > $\text{Mod}_{\text{Ability/Permission}}$'. For our purposes, we can conflate these as in (11).

Furthermore, Warner shows that epistemic interpretations of pre-modals emerge in ME (see also Lightfoot 1979, Chapter 1; Roberts 1985). We can interpret this as a further reanalysis of (some) pre-modals as being merged in the $\text{Mood}_{\text{Epistemic}}$ position in (11). This idea has the consequence that, if epistemic modals are merged higher than T(Past), they are unable to have the features associated with this position. We take this to mean that they are opaque to the usual past/non-past relation, a feature of the developing epistemic modals which has often been commented on (see Lightfoot 1979, Chapter 1; Roberts 1985; Warner 1993:148–150).

The idea that certain modals may have been directly merged in the functional structure is also consistent with the sporadic evidence, briefly alluded to above, that they may have been restructuring verbs. Cinque (2001) has argued that Italian clause-union constructions are indeed monoclausal, with the restructuring verbs actually merged in the functional structure. Put simply, Cinque's (2001) proposal is that restructuring verbs are functional heads. If so, then the indications that pre-modals may have been restructuring verbs are consistent with the proposal that they were merged directly into the functional system.

The tendency for certain modals to be directly merged in the relevant functional positions in (11) became categorical after the loss of infinitival morphology for the reasons given above. Note also that the argument given above that infinitival morphology provided a crucial cue for a biclausal structure is unaffected by the adoption of an elaborated clause structure, as long as we assume that infinitival morphology indicates the presence of a functional structure. In cases like those just described, the interpretation of the modal allows for the postulation of a monoclausal structure with the modal – necessarily finite – higher than the position of infinitive morphology. On other interpretations, in particular those where the modal had argument structure, the infinitival morphology provided the cue for a biclausal structure, exactly as described above. As we also mentioned above, the full syntactic effects of the reanalysis of modals were not apparent until after the loss of V movement into the functional system (which we can still tentatively identify as movement to T(Past)).

The above treatment of English modals has interesting parallels in other languages. What we observe in many languages is evidence for grammaticalization of individual modals, although the existence of a morphosyntactically distinct

class of the NE type is not attested elsewhere in Germanic or Romance. The reason for this, as we mentioned above, is that all the Germanic and Romance languages have infinitival morphology and so the reanalysis of (6) as (8) was not possible. This extends to the Mainland Scandinavian languages, which, as is well known, lack V-movement in non-V2 clauses (see Platzack 1987, Vikner 1995). As Benincà and Poletto (1997) show, the Italian deontic modal *bisogna* has some properties that are reminiscent of those of the English (pre-)modals: it lacks non-finite forms, personal forms and a range of tenses (see also the discussion on Greek *thelo* in section 1.3).[7] It also fails to host clitics and cannot have a subject (Denison (1990) and Warner (1993) argue that this last is also true of some English pre-modals). Benincà and Poletto (1997) also show that one version of *tocar* (literally 'touch') in the Veneto dialects of Padua and Venice is very similar to Standard Italian *bisogna*. Vikner (1988) shows that in Danish, epistemic modals cannot be non-finite, and van Kemenade (1985) argues the same for Dutch (see also Roussou (1999) for the non-availability of +past tense forms on epistemic modals in Greek). In the next section, we shall see some reason to think that Late Latin *habere* was also similar. If the above proposals are correct, all of these are instances of the grammaticalization of certain modal verbs on certain interpretations.[8] The full categorial split between modals and main verbs that we observe in English, from the sixteenth century on, is not found in these languages because infinitival morphology is retained.

Having presented the properties of English modals, their reanalysis, and the triggers for the change under consideration, let us go back to the structure in (11). In Cinque's (1999, 2001) terms, there is a distinct functional head for each reading associated with modal verbs. According to his analysis, epistemic modals should occur higher up in the clause than root modals. The clear advantage of

7. It has been pointed out to us that not all native speakers share the judgements provided by Benincà and Poletto (1997), and therefore the relevant restriction must be a temporal/aspectual one. Notice that if this observation is correct, it is still consistent with the present analysis, given that aspectual positions appear low in the functional hierarchy.
8. It may seem strange to propose that epistemic modals in Danish occupy a very high functional position like $\text{Mod}_{\text{Epistemic}}$ when the evidence is that all verbs, including modals, occupy just two positions in this language: the V2 position (presumably C) and what appears to be the base V position (see Vikner 1995 and the references given there). The problem really concerns associating the epistemic interpretation with the low position. This problem is just an instance of the general problem that arises in Mainland Scandinavian languages (and to some extent in English) of associating functional information (at the very least Tense) with the *in-situ* verb, and as such is not created by assuming the Cinque (1999, 2001) hierarchy. Whatever technical device we postulate to associate Tense with the *in-situ* verb (affix-hopping, chain-formation, LF movement, etc.) can be exploited to associate an epistemic modal with its functional position. See also the discussion that follows in the text.

his analysis is that it allows a single lexical item to receive different interpretations by simply assuming different positions in the functional structure. 'Lexical splits', then, can be simply derived syntactically, at least in the cases under consideration here. At the same time, his system turns out to be too powerful as there seems to be no limit on the number of functional heads postulated: in principle there could be a different head for each possible interpretation. There are cases though (as we will see in section 1.3 as well), where certain readings derive as a combination of other properties in the clause structure. Consider, for example, epistemic versus dynamic modals and the potential problem raised above, namely that the latter seem to have some sort of argument structure. In Cinque's (1999) system, though, the positions where arguments are structurally licensed are not represented. On this basis, it is not very obvious how certain interpretations interact with the presence versus absence of argument structure. Furthermore, as we saw in our discussion of English modals and we will see in the following two sections as well, a crucial factor in the reanalysis is the loss of agreement marking. If agreement marking correlates with argument structure, then we see once more that Cinque's structure fails to capture the correlation between the loss of inflectional properties and categorial reanalysis.

Despite the above problems we would like to maintain the spirit of Cinque's (1999, 2001) approach and maintain that (a) there is a structural correlation between a lexical item and the interpretation it receives, and (b) this correlation targets different heads in the clausal structure. However, it is possible to express these points by adopting a more conservative structure, such as the one in (12):

(12) $[_{TP}\ T\ [_{vP}\ v\ [_{VP}\ V]]]$

According to standard assumptions, lexical verbs are merged in V and they move to v and T (although this is subject to parametric variation – see Pollock (1989)). The VP shell in (12) also corresponds to the expression of the thematic structure, as argued by Hale and Keyser (1993). The lower VP determines the thematic role typically associated with objects, while the higher one determines that of subjects. Suppose, then, that dynamic modals, which seem to be a cross between lexical verbs and modals, are merged not in V but in v. If this is correct, we predict that they can participate in the determination of argument structure, and more precisely that of the subject. This is consistent with the fact that these verbs are subject oriented. This alternative approach also captures Zubizarreta's (1982) claim that these verbs assign 'adjunct' theta-roles. On the other hand, epistemic modals, which have no argument structure at all, are directly merged in T.

The tripartite system in (12) then correctly predicts that we can distinguish three types of verbs: epistemic modals, dynamic modals and lexical verbs. It

furthermore provides a clear correlation between the different positions and the absence or presence of argument structure. In languages like English where the lexical V can appear in the bare form, we predict that the structure in (12) straightforwardly captures the monoclausal structure of the modal+V sequence. We will see how this structure can account for the other changes discussed in the following two sections. The reanalysis of modals can be seen as involving two steps: (i) direct merge at v (with subsequent movement to T), and (ii) direct merge at T. When the modal is merged in v then it is still able to participate in argument structure and show T distinctions, and at the same time it is not regarded as a lexical V. If, on the other hand, the modal is merged in T directly, then it has no argument structure and doesn't show regular tense distinctions (it's just finite), yielding the interpretation typical of epistemic modals.

In this section, we have considered the development of the English modals, in the light of three independent changes. The first involves the reanalysis of some modals, and is attested in various languages as mentioned in the preceding paragraph. The second involves the loss of the infinitival morphology *-en* in English, which led to reanalysis of all modals. We proposed that the categorial split that took place early in the sixteenth century was triggered by the loss of infinitival morphology. This is also why this split is not found in other closely related languages. The third corresponds to the loss of V-to-T movement, which in combination with the loss of infinitival morphology gave rise to reanalysis of biclausal structures with modals to monoclausal ones. Notice that the last two are specific to English. Loss of V raising is found in Scandinavian as well, but not in association with loss of infinitival endings. It is the combination of all these three factors that led to the creation of a distinct class of modals in English. We have shown that this involved categorial reanalysis of the pre-modals as exponents of the T head (or v) in the functional system; in this respect, this change corresponds to our general characterization of grammaticalization as loss of movement.

2.2 *Romance futures*

Another very well-known case of grammaticalization is the development of the future and conditional tenses of most of the modern Romance languages (cf. Fleischman 1982, Pinkster 1987, Hopper & Traugott 1993:42–44, Roberts 1993b).[9] Traditional manuals of Romance philology, such as Bourciez (1967), Tekavčić (1980), describe these forms as originating in a periphrastic construction in Latin formed by an infinitive followed by *habere* (to have). For example,

9. In fact, not all of them, since southern Italian dialects and Sardinian do not have future tenses.

the future tense of nearly all Modern French verbs quite transparently shows this development. Compare the endings attached to the infinitive of *chanter* below, forming the future, with the present tense of *avoir* (to have):

(13) Future: chanter-*ai*, chanter-*as*, chanter-*a*, chanter-*ons*, chanter-*ez*, chanter-*ont*
avoir: ai, as, a, avons, avez, ont

The full lexical Latin verb *habere* was reanalysed as the future/conditional ending in the modern Romance languages in three stages. First, *habere* was reanalysed as a future auxiliary comparable to *will* in Modern English (cf. Fleischman 1982:60–66 on the relation between conditional and future in connection with *habeo*). This was a change from no realization of the future/conditional by a functional head to realization by an overt free morpheme, that is, Move > Merge (following Benveniste (1968), we take it that the Classical Latin future forms, e.g. *amabo* 'I will love' and *dicam* 'I will say', were being replaced by the periphrastic ones. One expects the synthetic forms to survive in later texts as well; cf. the discussion regarding the Greek future in the following section). The change affecting *habere* is extremely similar to the one involving English modals discussed in the previous section, except that it apparently involved only *habere*.[10] Second, the auxiliary *habere*, an autonomous word, was reanalysed as a syntactic affix. This is presumably a change from Merge to Move+Merge. The first change arguably took place in the third century, according to Benveniste (1968) (see below). The second change may be a direct reflex of the first (Fleischman 1982). The third change was the reanalysis of the syntactic affix as a lexical affix, that is, a feature of V, and the corresponding reintroduction of V-to-T movement in futures and conditionals. This change took place almost immediately after the earlier ones in French (Languedoïl), but slightly later in Occitan, Catalan and Northern Italian. It took place considerably later in Ibero-Romance (outside Catalan), as the evidence from clitic placement (which we will briefly go into) shows, and in fact may not have yet happened in conservative varieties of contemporary European Portuguese.

10. Actually, there is sporadic evidence for a similar reanalysis of *debere* (to have to), which is the future auxiliary in Logudorese Sardinian, and *velle* (to want), which became the future auxiliary of Rumanian (Fleischman 1982:114). Presumably, these changes took place at different times and in different places in the Latin/Romance-speaking area. The case of *debeo* is more interesting; presumably this verb comes under (17) in virtue of its form (2nd conjugation) and its meaning (stative, meaning 'to owe'). The development in Logudorese Sardinian is therefore expected, although we have nothing to say about why this development was not more widespread. The development of *velle* may be due to Greek influence (see next section), or may fall under (17) below since in Vulgar Latin *velle* became *volere* by analogy (J. C. Smith, personal communication).

The three changes can be roughly schematized as in (14), illustrated with the Italian future form *amerò* (cf. Roberts 1993b):

(14) a. $[_{TP}[_{VP}\ [_{XP}\ \text{amare}]\ t_{\text{habeo}}\ [_{T}\ \text{habeo}]]] > [_{TP}\ [_{XP}\ \text{amare}]\ [_{T}\ \text{habeo}]]$
b. $[_{TP}\ [_{XP}\ \text{amare}]\ [_{T}\ \text{habeo}]] > [_{TP}\ [_{XP}\ t_{\text{infin}}]\ [_{T}\ \text{amar} + \text{aio}]]$
c. $[_{TP}\ [_{T}\ \text{amar} + \text{aio}]\ [_{VP}\ t_{\text{infin}}]] > [_{TP}\ [_{T}\ \text{amer+ò}]\ [_{VP}\ t_{V+fut}]]$

We observe that, with the single difference of the relative order of T and its complement, the reanalysis shown in (14a) is the same as that relating (6) and (8) in the previous section – on the question of 'T-final' order in the light of Kayne (1994), see below. This is a further case, comparable to the ME modals and Italian *bisogna*, of 'leakage' of verbs with certain interpretations into the functional system. We can sharpen the parallel with the cases discussed in the previous section by deducing a certain morphological defectivity of the reanalysed *habere*. Although this verb had the full range of Latin tenses, voices and moods, only the present and imperfect (Gallo- and Ibero-Romance) or perfect (Italian) indicative active forms were reanalysed as futures and conditionals respectively. Thus we do not find future participles based on a Latin infinitive plus non-finite form of *habere* (and this despite the fact that Classical Latin had future participles, which, like the Latin future tenses, are entirely lost in Romance).[11] The absence of future participles and the like in Romance suggests that only a relatively small number of finite forms of *habere* were reanalysed. In other words, reanalysed *habere* was morphologically defective in a way which is directly comparable to the ME pre-modals discussed by Warner (1983, 1993) and *bisogna* as discussed by Benincà and Poletto (1997) (see previous section).

In Classical Latin, *habere* was a full verb with the core meaning 'to own' or 'to possess'. The following is an example of *habere* with a complement containing an infinitive where it is clear that *habere* is functioning as a verb of possession:

(15) De re publica nihil habeo ad te scribere.
Of thing public nothing I-have to you to-write
'I have nothing to write to you about the republic.' (Cicero; cited in Tekavčić 1980)

11. There seem to be some counter-examples to the claim that we don't find non-finite forms of *habere* with the infinitive. Fleischman (1982:55) gives the following example:

(i) tamquam ovis ad victimam adduci habens (Tertullian)
'like a lamb about to be taken to slaughter'

In this construction *habere* occurs with the passive infinitive *adduci*. As Fleischman points out this construction is used in the absence of a future passive participle. Crucially, though, the use of the participial auxiliary never gained much ground (Bassols 1948:305, cited in Fleischman 1982:55). We should perhaps relax our statement in the text, and assume that at some point in its development *habeo* could have non-finite forms, which it then lost.

In this kind of example, there is no reason to treat *habere* any differently from a standard transitive verb: it is a V with a DP complement (in the above example, this DP has a fairly complex internal structure) to which it assigns a theta-role. According to Sihler (1995:497) *habeo* formed an agentive-stative doublet with *capio* (take), cf. *pendo* (suspend) versus *pendeo* (I am hanging). We can interpret this to mean that *habeo* had an argument structure, although presumably one lacking an Agent thematic role. (Sihler (1995:531) points out that the stative verbs in these doublets were all second conjugation; on the potential significance of this observation, see below.)

The reanalysis of future/conditional *habere* gave rise to a 'lexical split', in that reflexes of *habere* in other contexts (essentially, where *habere* was not adjacent to an infinitive, see below) retain the possessive and related senses that are found in Classical Latin. This is true in Modern French and Italian, but possessive *habere* has been lost in contemporary Spanish and Portuguese. The reflexes of *habere* have also arguably been grammaticalized in other ways in other contexts: in existential constructions and in perfect tenses, although here again Portuguese has largely lost *habere*. Both possessive and perfect reflexes of *habere* have a full range of forms, including non-finite forms. Again, this indirectly points to an early morphological defectivity of future/conditional *habere*.

A further parallel with the ME pre-modals arises from observations made by Benveniste (1968). Benveniste clarifies a number of aspects of the developments in Late Latin. He points out that the periphrastic construction infinitive + *habere* originates with 'Christian writers and theologians starting with Tertullian', that is, early in the third century AD. The 'overwhelming majority of examples', according to Benveniste, indicate that the periphrasis involved a passive infinitive, as in the following (Benveniste 1968:90):

(16) in nationibus a quibus magis suscipi habebat.
among nations-abl by which-abl most to-be-accepted had
'Among nations by which the most was to be accepted.'

The periphrasis 'acts as the equivalent of the future passive participle' and 'served to indicate the predestination of an object to follow a certain course of events'. It seems clear that *habere* here has a (deontic) modal interpretation that essentially involves the notion of futurity. The 'thematic' interpretation of possession seems to be absent. In marking purely temporal content, *habere* is close to an auxiliary like Modern English *will*. In these constructions at least, then, *habere* had a defective argument structure from quite an early time (i.e. since it started being used as a modal). At this stage, it may be that *habere* could be merged under T in the structure in (12) (or $\text{Mod}_{\text{Necessity}}$ in terms of Cinque's (1999) substructure in (11)).

We can thus tentatively apply Warner's generalization for the ME pre-modals, given in (10), to *habere*, as follows:

(17) 2nd conjugation stative verbs subcategorized for an (passive) infinitive which denote necessity, obligation and related notions of futurity are finite only.

Since *habere* was the only second-conjugation stative to have the relevant semantic property, (17) singles out just this verb (*debeo* (owe) may have been another candidate, see note 10; other members of this class mentioned by Sihler (1995:531) include, in addition to those mentioned above, *video* (see), *taceo* (be silent), *sedeo* (be seated) and *iaceo* (lie down)). We can account for (17) exactly as we accounted for (10) in the previous section. If *habere* is inserted directly in T, it is unable to be licensed as non-finite, as already mentioned at the end of the previous section for English modals. Note that the generalization in (17) is not meant to explain the change that took place, but simply to show the parallelism with the English modals. The crucial point made by (10) and (17) is that a certain number of verbs form a group on the basis of their formal, and not their morphosemantic' properties.

Benveniste shows that between the third and the sixth century AD the periphrasis spread to a wider range of verbs and contexts: active intransitives and deponents. By the end of the Imperial period the periphrasis clearly had a straightforward future meaning (cf. Benveniste 1968, Bourciez 1967, Tekavčić 1980). The following example, from Tekavčić (1980:237), is a seventh-century case of clearly temporal *habere*:

(18) et quod sum, essere abetis
and that I-am, to-be habere-2pl
'And what I am, you will be.' (7th-century inscription)

The spelling has been Latinized in this example, and it is likely that it is somewhat distant from the contemporary pronunciation. The following is generally said to be the first example of the Romance synthetic future, and gives a clearer idea of the pronunciation:

(19) Iustinianus dicebat: 'Daras'.
Iustinianus said Give + habere + 2sg
'Iustinianus said: You will give.' (*Fredegario*, 7th century)

Here we see that the second singular of *habere* is reduced to *-as*. Tekavčić (1980:236) gives the forms of *habere* in this context as:

(20) *a(i)o, as, a(t), (av)emo, (av)ete(s), an(t)*

These forms, particularly the elimination of the stem *av-* in the first and second plural, indicate that future *habere* was aprosodic at this time. Whatever its phonological status, it seems clear that, by this time, *habere* was able to be merged in T.

As in the case of the English modals, we can see why the change in (14a) took place (we will discuss (14c), and why this change did **not** take place in English, directly). Considerations of complexity led the learner to select grammars that reduced the complexity of the representations that covered infinitive + *habere*. The older representation involved a V position in which *habere* was generated and a head-movement relation between *habere* and T, as shown in (14a) above. The innovative representation involved direct generation of *habere* in the functional system. With each reanalysis, a structure created by adjoining one head to another is eliminated. Again, as with the English modals, there is evidence for incremental reanalysis 'upwards' through the functional system. Of course, there are also significant differences between the English reanalysis discussed in the previous section and the Latin/Romance one. First, infinitive morphology was not lost in Latin/Romance (it still survives today). Second, *habere* became an affix, while the English modals remain morphologically free elements. We will now discuss these points in turn.

Regarding infinitival morphology, we mentioned in the previous section that the reanalysis of individual lexical verbs as functional heads restricted to finite forms and occupying 'high' functional positions is consistent with the retention of infinitival morphology. Indeed, this was exactly the situation in ME (and is the situation in contemporary Italian complements to *bisogna*). As we saw in the previous section, the loss of infinitival morphology in sixteenth-century English forced a split between auxiliaries and main verbs, but no such split has ever been forced in Latin or Romance, precisely because infinitival morphology has been retained in these languages.

The second point, namely the reduction of *habere* to an affix, requires us to take a closer account of Latin word order. Although Latin word order was rather free, there is a general consensus that the unmarked order was OV. This in turn implies that auxiliaries followed main verbs, following standard typological generalizations (see Greenberg 1966, Hawkins 1983, Croft 1990). In fact, we observe this order in relation to the infinitive and the auxiliary *habere* in examples such as (15)–(16), and (18)–(19) above. In order to account for similar orders in German, Cinque (1999:58), following Kayne (1994:141, n. 15), Zwart (1993:334f.) and Haegeman (1999), proposes that V may raise from VP and the remnant VP be fronted to the specifier of a functional head. In

line with these general ideas, we might propose a structure like the following for the relevant subparts of an example like (16):

(21) $[[_{VP}$ magis $t_V] \ldots [_T [_v$ suscipi$] [_T$ habebat$]]]$

Under this analysis, the head occupied by the auxiliary in fact has the Move+Merge option. We see that this option is characteristic of 'head-final' systems, as they are analysed in terms of Kayne's (1994) proposals. If it is true that $F^*_{Move+Merge}$ arises only in 'head-final' systems and only in this way, and if, as seems natural, this option is the principal means by which new affixes are created, then we can understand the general preference for suffixal morphology in the world's languages in terms of Kayne's proposal that adjunction is always to the left of the host.

In fact, there is an alternative way of capturing the V-final order in (21), while maintaining the results concerning 'head-final' systems. In particular, it is possible to argue that what occurs in T in (21) is actually the infinitive *suscipi*, while *habebat* occurs in a head adjacent to T, as shown in (21'):

(21') $[[_{VP}$ magis $t_V] \ldots [_T$ suscipi$] [_v$ habebat$]]$

In terms of this analysis the VP-internal material, such as *magis*, undergoes remnant VP movement, while V moves to T, bypassing the auxiliary. Whether it is actually the VP that remnant-moves, or simply the DP only that scrambles to a position above T, it's actually not easy to tell and to some extent not relevant to our present discussion. What is relevant though is movement of the infinitive to T over the auxiliary in v. Notice that this latter kind of movement makes this construction similar to other cases of Long Head Movement (LHM), where movement of V crosses over the auxiliary (cf. Lema and Rivero 1991, Roberts 1994). If this is correct, then we don't have an instance of $F^*_{Move+Merge}$ in (21), but F^*_{Move} for two distinct adjacent heads. Once the lower head becomes phonologically weak it reduces to a suffix and becomes reanalysed as a lexical suffix, that is, as part of the higher head. Of course, this is possible if there is strict adjacency between the two heads, that is, no material can intervene between the two, which is in fact the case as noted by Benveniste (1968). The correlation of OV with rich inflectional morphology still holds. There seem to be two reasons why this analysis is preferable. First, it gives us a strong correlation between the morphological and syntactic structure of future-marked verbs. Consider, for example, the French future *chanterai* or the Italian *amerò*. The original infinitival marker *-r* forms the future ending, while *-ai* and *-ò*, the *habere* residues, have reduced to a bundle of agreement features (essentially an agreement affix). This is illustrated in (22):

(22)

chante	-r	-ai	
ama	-r	-ò	
V	T	Agreement	

Given that V is in T we can capture the fact that it is *-r* in the verbal stem that marks future Tense in a straightforward way.

The second piece of evidence for this approach comes from those cases where clitics can intervene between the verb and *habeo*. There is evidence for this in all Romance varieties other than French, and particularly in Ibero-Romance. Here it appears that the medieval reflexes of *habere* were clitics rather than affixes. The evidence comes from the phenomenon of 'mesocliticization'. This term refers to the order *Infinitive-clitic-AIO*, found in Old Spanish and European Portuguese. An Old Spanish example of this is illustrated in (23) (see also Rivero (1997) for a discussion of the cliticization patterns in Old Spanish):

(23)

dezir	lo	hedes	al	rey?	
tell	it	you.will	to	the king	
'Will you tell it to the king?'					(Zif 124: Lema & Rivero 1991)

We can account for these orders if we assume that *aio* was a clitic auxiliary in these languages. Let us assume that, as a clitic, it was subject to the general medieval Romance ban on clitic-first orders (the Tobler-Mussafia law, cf. Benincà 1995). As a finite verb, pronominal clitics were proclitic to it. Mesocliticization is derived once we assume that non-finite V moves to the immediately higher position, as an instance of LHM, similar to that in (21'). In other contexts, the clitic precedes the sequence infinitive + *aio*, as in wh-questions below (Lema & Rivero 1991:250):

(24)

A	quien	nos	dar	edes	por	cabdiellos?
to	who	us	give	you.will	for	leader
'Who will you give us as leader?'						

The clitic precedes the infinitive just where it would not thereby come first in the string. Thus it appears that in these varieties, as long as mesoclisis remains, *habere* is not reduced to an affix, but remains a clitic auxiliary. The reanalysis as an affix is presumably blocked by the cliticization evidence (see also Wanner 1987) (we will come back to this point when we discuss the Greek future particle *tha* in the following section).

This analysis of mesoclisis links it to the ban on first-position clitics. Lema and Rivero (1991) show that the ban on first-position clitics and mesoclisis are lost together in Spanish in approximately the sixteenth century, confirming the idea that the phenomena are related. The ban on first-position clitics is found in

all the Romance languages at some stage in their history, and, indeed, mesoclisis is attested, albeit sporadically, in all Old Romance varieties except French (i.e. Languedoïl). Moreover, both (a form of) the ban on first-position clitics and mesoclisis are still found in contemporary European Portuguese (see Madeira 1995).

The current analysis allows us to account for the development of *habere* as an affix, and for why no comparable development affected the English modals. This difference follows from the simple fact that Latin was OV, while sixteenth-century English was not. Hence *habere* triggered LHM, as shown in (21'), giving rise to an Aux-final order, while English modals never did. Since ENE was VO, the modals did not develop into affixes.

We can now understand the third change in terms of the change from OV to VO word order that took place between Latin and Romance. The OV to VO shift was accompanied by a shift from superficial VP-Aux order to Aux-VP order. The obvious analysis of this in terms of what we have just said about OV languages is to say that leftward movement of VP and the associated V movement were lost (see Roberts 1997a for a sketch of how these changes may have taken place in Early ME; there a connection is made to the loss of morphological nominative-accusative distinctions – such distinctions were also lost in the transition from Latin to Romance, see also the discussion in Chapter 4, section 4.1). This change is of course a simplification, since movement dependencies are lost. In other words, when VP-Aux orders were lost as a part of the general word-order change, the sequence infinitive + *habere* was reanalysed as a single word and so the auxiliary became a lexical affix. This is apparently what happened in French, while the situation in other varieties is slightly more complex owing to the existence of mesocliticization, as we saw above. Fleischman (1982:121) also links the development of *habere* as an affix to the earlier OV word order and also relates the fact that perfect *have* has not developed into an affix in Romance to the fact that this construction developed later; note that the latter point follows on the present account for the same reason as the fact that the English modals have not developed into affixes.

A potential problem arises with respect to the history of the perfect *habeo*. Nigel Vincent (personal communication) has pointed out that there seems to be evidence that perfect *habeo* grammaticalized earlier (by the first century BC) than the future construction with *habeo*. Assuming that Latin was OV at this point, the question is why perfect *habeo* was not affected, that is, why *habeo* in this case did not reduce to an affix. Consider first the OV order and its significance for our analysis. Notice that our account of the reanalysis of future *habeo* to a suffix holds as long as it is empirically supported, that is, as long as we find

data where the infinitive precedes the auxiliary *habeo* and there can be no intervening material (strict adjacency). The data discussed in the literature seem to point in this direction. Whether the infinitive + *habeo* pattern reduces to a more general OV ('head-final') pattern can indeed be left open (see also Nocentini forthcoming). In other words, we can still maintain our analysis, assuming that the V–Aux order is an instance of LHM, while relaxing the typological correlation with OV. It is perhaps worth pointing out in this connection that, in a Kaynian antisymmetrical framework of the kind we are adopting here, the notion 'OV language' has no theoretical status; systems are OV to a greater or lesser extent depending on the greater or lesser incidence of XP movement to the left. Therefore it is quite unproblematic to assume that Latin may not have been 'fully OV' at the same time as having the leftward-movement operations which gave rise to the crucial infinitive–*habere* sequences. The second point concerns what prevented perfect *habeo* from following this pattern, especially if perfect *habeo* grammaticalized earlier than its future counterpart. We would like to suggest tentatively that the presence of *habeo* in perfect constructions at this early stage does not necessarily imply that *habeo* is an auxiliary. Instead we could assume that it is a lexical verb taking a participial (small clause) complement.

According to Harris and Campbell (1995:183) the *habeo* + (passive) participle construction in Latin (and early French) has the following distinctive properties: '(i) the possibility of distinct subjects in the two clauses; (ii) agreement; and (iii) word order'. This is illustrated with the following two examples from Latin and early French (cited in Harris & Campbell 1995:182–183):

(25) a. in ea provincia pecunias magnas collocatas habent.
in this province capital great invested have-3pl
'They have great capital invested in that province'

b. et [chis empereres] avoit letres seur lui escrites qui . . .
and this emperor has-3sg letters on him written which . . .
'and this emperor has letters written on him, which [say] . . .'

The examples in (25a) (from Cicero, cited by Vincent 1982:82) and (25b) illustrate the above properties. According to the above authors the Latin perfect construction can be analysed as a biclausal one. In our terms (but not in Harris & Campbell's) we could say that it has the following structure: *they have [great capital invested]*. The presence of systematic agreement on the participle can be taken as evidence for this construction. The situation changes in Old French, where participial agreement depends on the position of the object, that is, whether it precedes or follows the participle, and also where SVO is the

established word order. At this point perfect *habeo* can be taken to function as an auxiliary, as also pointed out by Harris and Campbell (1995). Whatever the exact details of the analysis, it allows us to assume that perfect *habeo* in (Late) Latin was not an auxiliary. Instead it seems that this grammaticalization took place later than that of future *habeo*. The chronological difference could be attributed to a number of reasons, one of which is the different complements found with perfect and future *habeo* (participle vs. infinitive respectively). If this is correct, then we can maintain the essence of our analysis and its possible correlation with word-order changes.

To conclude, we can see that the changes which led to the creation of the Romance futures were partly rather similar to those which affected the English modals, but that independent differences led to a different development. The first change, the development of *habere* into a future/modal auxiliary, was just like the sporadic grammaticalization of the ME premodals, and motivated by exactly the same factors: the possibility of interpreting *habere* as an element with defective (or no) argument structure and the consequent possibility of economizing movement by merging it directly as a functional head. The change of *habere* into an element triggering head movement followed from this change combined with the OV typology of Latin. Finally, *habere* became a pure affix (a feature of V triggering raising to a functional position) when certain independent developments in the clitic system took place (the loss of the ban on initial clitics).

In the following section, we will consider the development of the future particle *tha* in Greek, and see how its development relates to that of the two other cases discussed so far.

2.3 *The Greek future*

The third case we will consider is the grammaticalization of the 'future' particle *tha* out of the volitional verb *thelo* (want) in Greek. Classical Greek had a synthetic future very much like Latin, while Modern Greek (MG) uses a periphrastic construction consisting of the particle *tha* and the verb. The loss of the synthetic future goes back to the *Koine* (i.e. the Greek of the Hellenistic and Roman period, third century BC–fourth century AD), when various phonological changes, which are already attested in the early Hellenistic period, affected the vowel system of Greek. As a result of these changes the morphological paradigms of future indicative and aorist (past tense) subjunctive basically became homophonous (cf. Browning 1983:25–26, Horrocks 1997:108ff). While the aorist subjunctive (which still stood in opposition with the aorist indicative)

was also used to express the future, a number of periphrastic constructions also emerged. These were formed with verbs like *ekho* (have), *mello* (be about), *thelo* (want), *opheilo* (owe) followed by the infinitive. Of those, *thelo* + infinitive became the main future expression in the Byzantine period, around the tenth century (Joseph 1990:116, Horrocks 1997:167), while *ekho* + infinitive was restricted to the expression of the perfect tenses, thus differing from the Romance case discussed in the previous section. The gradual loss of the infinitive and its replacement by a finite complement also gave rise to an alternative construction with *thelo*, namely one that has a *(h)ina*-complement (*hina* is the predecessor of the subjunctive particle *na*).[12] It is a rather standard assumption in the literature that *tha* developed out of some form of *thelo* (*thelei* > *thel'* > *the*) plus *na* (cf. Jannaris 1897, Chatzidakis 1905, Meillet 1912, Bânescu 1915, Joseph 1983, 1990, Horrocks 1997, among others). Thus *tha* has been analysed as a typical case of grammaticalization: a lexical verb reduces to a grammatical marker (cf. Meillet 1912, Hopper & Traugott 1993, Bybee *et al.* 1994, McMahon 1994, Harris & Campbell 1995, Tsangalidis 1999).

Before we discuss the grammaticalization of *tha* in more detail, it is worth briefly describing its distribution in MG. Although *tha* is usually called the 'future' marker, it is not just restricted to a future context, but it gives rise to a number of modal readings, depending on the tense (+/−past) and aspectual (+/−perfective) properties of the verb. The future reading clearly arises when the verb is –past, +perfective, as shown in (26) below (see Tsangalidis 1999:212); prt = particle.

(26) a. Tha *egrapse* to grama. (egrapse = +past, +perfective)
prt wrote-3sg the letter
'He would/must have written the letter.'

b. Tha *egrafe* to grama. (egrafe = +past, −perfective)
prt wrote-3sg the letter
'He was supposed to be writing the letter.'

c. Tha *grapsi* to grama. (grapsi = −past, +perfective)
prt write-3sg the letter
'He will write the letter.'

d. Tha *grafi* to grama. (grafi = −past, −perfective)
prt write-3sg the letter
'He must be writing the letter.'

If *tha* occurs with the –past, −perfect form of the verb (*grafi*) the epistemic reading is the preferred one, although the future interpretation is also possible

12. On the loss of the infinitive see Joseph (1983). For a more recent discussion of *na* see Philippaki-Warburton and Spyropoulos (2000) and the discussion in Chapter 3, section 3.1.

given the right context: for example, by adding the adverbial expression *tomorrow*. When *tha* occurs with +past, +perfective (*egrapse*), or +past, −perfective (*egrafe*) forms of the verb, the future reading is blocked, while an epistemic and/or counterfactual reading is possible. On this basis we analyse *tha* as a modal particle instead of a future (tense) marker. We further assume that *tha* is the head of a Modal projection which follows NegP (with negator *dhen*), as shown in (27) (cf. Rivero 1994, Drachman 1994, Roussou 2000) (we will elaborate on the structure in (27) in Chapter 3, section 3.1):

(27) $[_{NegP}$ Neg $[_{MP}$ tha $[_{TP}$ T . . .]]]

On this assumption, the grammaticalization of *tha* involves reanalysis from a lexical V to a modal particle high up in the functional structure.

Notice that *thelo* in MG is a lexical verb of volition, taking a *na*-complement, unlike its Classical Greek counterpart that subcategorized for an infinitive, as shown in (28a) and (28b) respectively:[13]

(28) a. Thelo na grafo.
want-1sg prt write-1sg
b. Thelo: graphein.
want-1sg write-inf.
'I want to write.'

Given that *thelo* remained as a volitional verb, there must have been a stage during which a 'lexical split' took place, giving rise to a lexical *thelo* and an auxiliary one which ultimately expressed futurity (cf. Pappas & Joseph 2001, see also Beths 1999 on English *dare*, and the discussion in the above sections). Thus the grammaticalization of *tha* involves three basic stages: lexical verb > auxiliary > particle. If *tha* is a Modal head, high up in the functional structure, and if auxiliaries are in T (cf. sections 2.1–2.2), then the grammaticalization of *tha* involves reanalysis from a lexical head (V) to a lower functional head, and then to a higher one. We will show that in its status as a particle, *tha* does not bear V features, thus differing in this respect from English modals, as well as the future affix in Romance. (In fact, as we will show in Chapter 3, section 3.1, *tha*, just like the subjunctive particle *na* are modal particles realizing M in the C system.) On the other hand, *thelo* as an auxiliary retained its V features. The loss of V features structurally corresponds to merger in M in the high functional field.

13. Notice that we use a different transliteration for the MG and CG data. Where MG has [f] (a labiodental fricative), CG had [p^h] (an aspirated labial stop). Furthermore, as the examples in (28) show CG has long vowels, while MG doesn't.

Let us now consider the changes that gave rise to *tha*. The discussion that follows is mainly based on Joseph (1983, 1990) and Pappas and Joseph (2001). One of the major syntactic changes attested in the *Koine* was the gradual loss of the infinitive. Infinitival complements were replaced by an *oti-* or *(h)ina*-clause, depending on the matrix predicate. Roughly speaking, verbs of saying, assertion, supposition and factives subcategorized for an *oti*-complement, while all the others took a *(h)ina*-complement. However, raising (*mello* 'be about', *opheilo* 'ought', etc.) and control (*tolmo* 'dare', *epithimo* 'desire/wish', etc.) predicates retained their infinitival complement. The verb *thelo* showed a double pattern: when the matrix and the embedded subject were coreferential (control), the infinitive was selected; when the two subjects were disjoint in reference, the *(h)ina*-complement with a verb in the subjunctive was selected. This is actually reminiscent of the situation we find in may modern Romance languages: control is compatible with infinitives, while a subjunctive complement requires disjoint reference (the obviation effect) (cf., for example, Picallo 1985, Kempchinsky 1986, Farkas 1992, among others). This systematic pattern broke down around the second century AD, allowing for coreference with an *(h)ina*-complement as well, as shown in (29) (from Joseph 1983:53).[14] In other words, the finite complement allowed for free reference, exactly as in MG:

(29) thelousin hoi Ioudaioi hina phoneusousin auton.
want-3pl the Jews-nom that kill-3pl him
'The Jews want to kill him.' (Act. Pil. 11.2.5)

We then see that infinitival complements were restricted to a small set of contexts which formed a proper subset of the *(h)ina*-complements. Some infinitival complements remained until the late Byzantine/medieval period (from the eleventh century onwards, extending up to the seventeenth century in some cases) (Joseph 1983:57). The future construction *thelo* + infinitive is a typical example of this. The availability of two distinct complements for *thelo* (infinitive vs. a finite clause) in the *Koine* period created the conditions for the two readings of *thelo*, that is, as an auxiliary and a lexical verb. In later stages (from the tenth century onwards), *thelo* + infinitive is mainly restricted to the future reading, while *thelo* + *na* expresses volition, exactly as in MG. Thus the lexical versus functional distinction of *thelo* is syntactically expressed on the complement (cf. Horrocks 1997:231, Tsangalidis 1999:151).

The sporadic use of the infinitive in the late Byzantine/medieval period is found with those predicates that took an infinitival complement in Hellenistic

14. In all the examples that come from Joseph's work we use his transliteration.

Greek. Interestingly, these verbs (e.g. *arkhomai* 'start', *thelo* 'want') are aspectuals or modals, and belong to the class of restructuring verbs:

(30) a. eis touto arksetai lalei (Morea 3824 (P), 13th century)
to this begin-3sg speak-inf
'At that he begins to speak.' (in Joseph 1983:58)
b. kathos to theleis mathei (Morea 1197)
as it want-2sg learn-inf
'As you will learn it . . .' (in Joseph 1983:64)
c. Gianni *lo* vuole fare.
John it want-3sg do
'John wants to do it.'

The example in (30b) with *thelo* shows clitic-climbing, a property typical of restructuring predicates also exemplified by the Italian (30c). Assuming that restructuring verbs trigger clause-union (see sections 2.1 and 2.2), (30a–b) are essentially monoclausal (cf. Joseph (1990, Chapter 5) for more arguments). In other words, we have reanalysis from a biclausal to a monoclausal structure, as in (31a). The structure in (31b) is the equivalent with a *na*-complement which does not show clitic-climbing and therefore no direct evidence for a monoclausal structure:

(31) a. $[_{TP}$ to theleis $[_{VP}$ t_v $[_{TP}$ mathei$]]] \rightarrow [_{TP}$ to theleis $[_{VP}$ mathei$]]$
b. $[_{TP}$ theleis $[_{VP}$ t_v $[_{CP}$ na to matheis$]]]$

(31a) is consistent with what we argued for regarding the English modals and the auxiliary *habere* in Post-Classical Latin in the previous sections (recall also that this is consistent with Cinque's (2001) approach to restructuring, and the modification we provided in section 2.1). On this basis, we can argue that auxiliary *thelo* is merged in a functional head above the VP. So at least in the early stages of its development as an auxiliary, we can assume that it merged directly in v, and from there moved to T. The reanalysis of *thelo* as a T element goes along with the availability of a +past tense specification (and absence of any possible non-finite forms). Indeed, combinations like *ithela grapsei* (wanted-1sg write-3sg) are possible, albeit with a counterfactual reading (Pappas 2001). Lexical *thelo*, on the other hand, is merged in V, has argument structure, takes a CP complement, and moves to T (via v). At this point, then, we have a change from Move to Merge as far as the auxiliary *thelo* is concerned.

The next obvious question concerns the trigger for the monoclausal reanalysis in (31a). Recall that in our discussion of the English modals, we argued that reanalysis is triggered by the loss of the infinitival ending *-en*. If the *thelo* construction is also reanalysed as monoclausal, then we expect to find a similar loss of

infinitival morphology. This indeed turns out to be the case. The table in (32) below shows the reduction of the inflectional paradigm of the infinitive in Medieval Greek (from the twelfth century onwards) compared to that in Classical Greek. The new system has a single ending *-ein* [in] for the active infinitive and a single ending *-the:(n)* [θin] for the medio-passive (from Joseph 1990:23):

(32) a. *Classical Greek*:

	Active	Passive	Middle
Present/future	*-ein*		*-sthai*
1st aorist	*-(s)ai*	*-the:nai*	*-sthai*
2nd aorist	*-ein*	*-e:nai*	*-sthai*
Perfect	*-enai*		*-sthai*

b. *Medieval Greek* (from 12th century onwards)

	Active	Medio-passive
Present/future	*-ei(n)*	*-the:(n)* (loss of final vowel by analogy)
1st aorist	*-(s)ei(n)*	*-the:(n)*
2nd aorist	*-ei(n)*	*-the:(n)*

Apart from being morphologically reduced, the crucial change in (32b) concerns the loss of the final *-n* (see (30a, b) and (31a)), which makes the infinitival ending *-ei* [i] homophonous with the third-person singular (−past) form, as shown in (33) below:

(33) thelei grafein → thelei grafei
want-3sg write-inf want-3sg write-inf/3sg?

While the loss of the infinitival morphology in English gave rise to a bare stem form, in Greek it made the infinitive formally identical to a finite form.

To this end, the loss of final *-n* not only removed the trigger for a biclausal structure, in the sense that there was no clear evidence for an infinitival T (i.e. for T^*_{Move}), but also gave rise to reanalysis (or perhaps 'misanalysis') of the infinitive to a finite form. The result of this change was 'agreement spreading' to all persons, as shown in the following examples (according to Bânescu 1915, cited in Joseph 1990:116 these forms are attested in texts of the fifteenth century):

(34) a. theloun armatosoun to koumounin (Makhairas 372, 1.22, 15th century)
want-3pl outfit-3pl the expedition
'They will outfit the expedition.'
b. 's to telos thelo sas to po (Bios Dem. 398, 16th century)
at the end want-1sg you it say-1sg
'In the end I will tell you it.' (in Joseph 1983:66)

c. dhe thes evris (Erotokritos A 1527, 17th century)
not want-2sg find-2sg
'You won't find.' (in Holton 1993:122)

d. kai panta thelo s' agapo (Gyp. I. 382 (A), 17th century)
and always want-1sg you love-1sg
'and I will always love you' (in Joseph 1990:136)

Thus *thelo* at this stage takes a complement that has a finite V but no subordinator (a VP perhaps), or a full CP introduced by *na* in its volitional reading, as shown in (35):

(35) a. thelo grafo
want-1sg write-1sg

b. thelo na grafo
want-1sg prt write-1sg

The question then is whether we can still maintain a monoclausal analysis for (34) and (35a).

At a first approximation, it seems that the only structural difference between (35a) and (35b) has to do with the presence of the particle *na* in the latter case. One possible analysis is to assume that *na* optionally deletes, yielding two different outputs at PF. Extending this analysis to the examples in (34), we would also say that what follows *thelo* is a CP with optional deletion of *na*. This would in turn imply that the monoclausal structure in (31a) goes back to being biclausal, as in (31b), albeit with C being optionally phonologically empty. However, this approach turns out to be problematic for a number of reasons. First, the two structures in (35) receive different interpretations. While (35a) is the future construction, (35b) is the volitional one, that is, in the former *thelo* does not have argument structure (it is an auxiliary), while in the latter it does (it is a lexical verb). Second, finite complements in Greek are always introduced by a subordinator (leaving aside some marked cases of *oti*-deletion in semi-direct speech in MG at least). In this respect the optional deletion of *na* in (35a) (and (34) for that matter) would come as a surprising exception, even more so as it would have to be restricted to the verb *thelo*. Third, although there appear to be two finite verbs in (35a), only one of them, namely *thelo*, can inflect for tense, as shown in (36) (there are no data in the literature where both forms inflect for past tense):

(36) a. k' *itheles* to 'kheis thamasma (Gyp. I.70, 17th century)
and wanted-2sg it have-2sg wonder
'And you would regard it a wonder.' (in Joseph 1990:135)

b. ithela t' agroike:so (Kats. *Thy*. V.316, 18th century)
wanted-1sg them hear-1sg
'I would hear them.' (in Joseph 1990:136)

As already mentioned, when *thelo* is +past, as in (36), it yields a counterfactual interpretation (cf. Pappas 2001). Apart from this, (36) and (34) are structurally alike (see the finite form of the main verb and the position of the clitic).

The double pattern illustrated in (36) is also attested in some Southern Italian dialects, for example Salentino (which has a very restricted use of infinitives) (Calabrese 1993:81–82):

(37) a. Voggyu (*ku*) kkattu nu milune
want-1sg (that) buy-1sg a melon
'I want to buy a melon.'

b. Voggyu *(*ku*) vvyeni kray.
Want-1sg (that) come-2sg tomorrow
'I want you to come tomorrow.'

Calabrese (1993) argues that *ku* (the equivalent of *na*) is optional when coreference is at stake, but obligatory with disjoint reference. In his analysis, (37a) is a case of optional *ku*-deletion. This construction is quite reminiscent of (35), given that when *ku* is absent the verb 'want' takes what looks like a finite V as its complement. As far as (35a) is concerned, we excluded the possibility of *na*-deletion. The question then is whether something similar also holds for (37a) in the absence of *ku*. In other words is it possible that (37a) also involves two different constructions, that is, one with a CP complement, and one with two finite Vs? Manzini and Savoia (forthcoming) argue that the *ku*-less example in (37a) is a serial verb construction ($V_1 + V_2$), thus a monoclausal structure. On the other hand, when *ku* is present there is a biclausal construction with a CP complement. The same account can extend to the Greek data in (35): when *na* is present *thelo* takes a CP complement, as in (31b). When *na* is absent, the auxiliary *thelo* occurs in a high functional head, while the main verb occurs either in V or in a lower functional head (presumably v).

Despite similarities, the Salentino data differ from the Greek ones in a number of ways. First, *want* in (37a) is interpreted as volitional irrespectively of the presence of *ku*, while in Greek we get two different readings depending on the presence of *na*. Second, while the *ku*-less construction in Salentino shows proclisis on the verb *want*, this is not the case in the Greek data under consideration. Going back to the data in (34b, d), as well as (36), we notice that the clitic follows *thelo*. Comparing these data to (30b), it is clear that there is a change in the position of the clitic, which now follows and does not precede *thelo* (see Joseph 1990, Chapter 5).

There are two ways to account for this change. The first option is somehow to link this change to agreement spreading. More precisely, in the 'agreement spreading' case, it is the lexical V and not *thelo* that carries the primary

agreement, which is in turn doubled by the auxiliary.[15] The verb *thelo* itself is merged in a high functional head that precedes the clitic position in the clause structure. The clitic, being an inflectional element, attaches to the verb that carries the main agreement, that is, the lexical verb. The alternative is to attribute this pattern to a change in the position of the clitic, bearing in mind that the clitic+V order (proclisis) of MG developed out of an enclitic one (see Horrocks 1997:210ff. for a discussion of cliticization in vernacular Medieval Greek, and Mackridge 1993).[16] Even if we take the latter approach though we would still have to assume that the verb *thelo* has moved to a higher functional position in the clause structure, so that it would precede the clitic. In other words, *thelo* further raised from T to C (or more precisely M), leaving the clitic in the I domain. The next stage of course would involve direct merger of *thelo* in M.

According to Joseph (1990), Pappas and Joseph (2001), the next developmental stage in the '*thelo grafo*' future involves reduction of agreement on *thelo* which surfaces as impersonal *thelei* (3rd singular) (attested in texts from the sixteenth century onwards):

(38)

a. [...] thelei sou dhosou ji andra sou (Panoria C' 245, 16th century)
want-3sg you give-3pl for husband yours
'[and] they'll give you as your husband' (in Markopoulou 2000)

b. ki emeis thelei ta kamome (Kats. *Thy.* II.322, 18th century)
and we want-3sg them do-1pl
'and we will do them' (in Joseph 1990:119)

Along with the *thelei*+V future, we also find the *the na* construction, which is actually attested in earlier texts, certainly from the fourteenth century onwards (see Jannaris 1897:558, Horrocks 1997:232):

(39)

a. kai plio dhe the na kartero (Erotokritos A 1231, 17th century)
and more not want prt wait-1sg
'and I won't wait any longer' (in Holton 1993:122)

b. autonos the na xanetai ston potho ojia mena (Erotokritos A 800,
he want prt lost-3sg in-the desire for me 17th century)
'he will be getting lost in the desire for me'

15. This could be analysed as an expletive-associate chain involving agreement affixes instead of DPs. The agreement on the main verb is 'argumental', while that on the auxiliary is 'expletive'. This is a possible analysis, bearing in mind that Greek has been a null-subject language throughout its history.
16. MG has proclisis, i.e. clitic + V (with the exception of imperatives and gerunds which show enclisis). The enclitic pattern is still attested in some Greek dialects, notably Cypriot Greek (see Rivero & Terzi 1995, Agouraki 2001, among others).

The standard idea is that *the* is a reduced form of impersonal *thelei* (*thelei* > *thel'* > *the*). If this is correct, then there must have been a stage where we find impersonal *thelei na* constructions. Although we have written records of *thelei*+V, this is not the case with *thelei na* (i.e. where *thelei* gives rise to a future interpretation). The question then is how we proceed from *thelo na* to *the na* (the latter being attested in the written records).[17] Notice that in the texts of the Cretan Renaissance (up to the seventeenth century), we find the *thelo*+V future (both with 'infinitival' V and agreement spreading), the *the na*+V future,[18] as well as some examples with *tha* (Holton 1993). The following example (from Chila-Markopoulou 2001:826) is revealing of this situation:

(40) enas mas *the na skotothei* ki o rigas tou *tha xasei.*
one ours will prt be-killed-3sg and the king his prt lose-3sg
'One of us will be killed and his king will lose.' (Erotokritos D', 1778–1780)

The Cretan comedies show an increased use of *tha* as well as examples of impersonal *thelei*+V. Holton's (1993) quantitative analysis shows (ignoring genre differences) that grammaticalization of *tha* is already completed in the late sixteenth century. Thus, if there was a *thelei na*+V future construction, with *thelei* being impersonal, that must have occurred much earlier.

Pappas and Joseph (2001) reconstruct the (impersonal) *thelei na*+V stage, as a necessary step for the reduction to *the na* > *tha*. However, this reconstruction is not directly supported by the data, in the sense that there are no written records of impersonal *thelei* taking a *na*-complement. The alternative would be to assume, following Horrocks (1997) (who follows Jannaris 1897), that the reduced form of *thelo*, namely *the*, was used to strengthen the future interpretation which was also conveyed by the *na*+V construction, that is, the reanalysed form of the subjunctive (but see Joseph 1990, Pappas & Joseph 2001 for counter-examples). Even if we accept this alternative, we still have to explain how we get *the*, or how *the* develops, in the future construction with *na*.

17. There may be some indirect evidence for this form that comes from the Cretan dialect, as in (i) (from Chila-Markopoulou 2001:827):

(i) all' as einai, na sas to po thelei ki afto.
but prt be-3sg, prt you it say-1sg want-3sg and this
'But let it be, I'll also tell you this.'

The other piece of evidence for impersonal *thelei* comes from examples like those in (38).

18. Holton (1993) points out that while the *thelo* + 'infinitive' future construction is restricted to the active voice in the epic poem of *Erotokritos*, the *the na* construction is not. He then suggests that the availability of the full set of voice and aspect distinctions in the *the na* construction, contributed to it being preferred over the *thelo* + infinitive future.

Let us start with *the* first. In connection to this, notice that the verb *thelo* in MG can take the following forms:

(41) a. thélo, théleis, thélei, théloume, thélete, théloun.
b. thélo, thés, thélei, théme, théte, thén(e).
'I want, you want, he/she wants, we want, you want, they want.'

The paradigm in (41) shows that the verb *thelo* has a series of reduced forms in second singular and the plural. In fact, a similar reduction is found for third singular in fast speech, especially when followed by *na*: [θeli] > [θelə], although this is not presented in the written language. (In certain dialects, the forms in (41b) can further reduce, with the elimination of the final unstressed vowel.) Bearing these observations in mind, we could say that the reduced *the na* form (for auxiliary *thelo*) was derived from the paradigm in (41b). As Pappas and Joseph (2001) point out, the residue of a final [l] is found in dialectal forms of *the na*, such as the Cypriot form *enna*. In this respect *the* can be to some extent supported synchronically.

Consider next the form *thelei* which in the written records only appears as an impersonal with the 'infinitival' complement, but not with the *na*-complement. We could, however, provide some synchronic evidence for the reconstructed *thelei na* sequence, based on the non-volitional uses of *thelo* in MG, as in the examples below:

(42) a. Ta rouxa theloun/*thelei plisimo.
the clothes want-3pl/want-3sg washing.
'The clothes need washing.'
b. Ta rouxa theloun/thelei na ta plinis kala.
the clothes want-3pl/want-3sg prt them wash-2sg well
'The clothes need to be washed well./ You ought to wash the clothes well.'

In (42) the verb *thelo* translates as 'need' (non-volitional). In (42a) it takes a nominal complement, which is actually a deverbal noun, whereas in (42b) it takes a *na*-complement. Notice that when the nominal complement is present, agreement is obligatory. However, with the CP complement, the verb can either agree in number with the subject, yielding *theloun* (the personal construction), or not, yielding *thelei* (the impersonal construction). There is actually a slight difference in meaning, as in the personal construction the sentence translates as '*x* needs to be washed', while in the impersonal one it translates as 'you/one ought to wash *x* well'. We will not discuss the examples in (42) in detail (but see Roussou 2002a). It is interesting to note that the constructions in (42b) are considered rather colloquial, but are nevertheless quite productive. For this reason, their absence from the written language can be expected. At the same time, the availability of non-volitional *thelo* in MG and in particular of impersonal *thelei* can be used as evidence for the existence of an intermediate *thelei na* stage which gave rise to the *the na* construction, on the basis of the phonological reduction described in (41).

Notice that non-volitional *thelo* in MG does not take the full set of inflections. For example, it is incompatible with the gerundive form, as well as with perfective aspect:

(43) a. *Thelondas plisimo/na ta plinis kala
needing washing/prt them wash-2sg well
b. *Ta rouxa thelisan plisimo /na ta plinis.
the clothes needed-perf-3pl washing/prt them wash-2sg
c. Thelo na me plinis.
want-1sg prt me wash-2sg
'I want/*need you to wash me.'

The gerundive (43a) and perfective (43b) forms are only available to volitional *thelo*. Furthermore, first- and second-person specification on *thelo* in the presence of a *na*-complement, necessarily triggers the volitional reading, and blocks the non-volitional one (and therefore can only be derived when the subject is animate, as volition is a property of animate entities. Notice that non-volitional *thelo* does not impose any animacy restrictions on its subject). We can then see that MG also shows a distinction between two readings of *thelo*, one of which is more grammaticalized than the other. So in present terms, volitional *thelo* is merged in V and moves to v and T, whereas non-volitional *thelo* is not merged in V (in some cases it can be merged in v, while in others, that is, when it has more of an epistemic reading, we could assume that it is merged in T).

What the preceding discussion shows is that although there may be no direct evidence for the presence of an impersonal *thelei* followed by *na* in its development as a future marker, we can reconstruct this stage by using indirect evidence from its non-volitional use in MG. On the basis of the discussion in this section we can summarize the changes in *thelo* + infinitive as in (44), and those involving *thelo* + *na* as in (45):

(44) a. *thelo* + infinitival V: agreement on *thelo* only
b. *thelo* + 'finite' V: agreement spreading
c. *thelo* + finite V: expletive agreement on *thelo*
d. *thelei* + finite V: impersonal (default agreement on) *thelei*

(45) a. *thelo* + *na*-clause: volitional *thelo*, referential agreement
b. *thelei* > *the* + *na*-clause: auxiliary, no agreement
c. *tha* + finite V: *tha* as a particle

Fully inflected *thelo* could be ambiguous between an auxiliary and a lexical verb. The ambiguity was partly resolved structurally by means of the complement clause: an infinitive for the former and a *na*-clause for the latter (cf. (44a) and (44b) respectively). *Thelo* with 'expletive' agreement in (44c) corresponds to auxiliary *thelo* only. Although it is morphologically identical to lexical *thelo* it forms a verbal complex with the finite verb which carries the primary agreement. This is supported by the position of the clitics, which at this stage mainly follow *thelo*. The next step in (44d) involves impersonal *thelei*: the verb in this case does not inflect for person (1st and 2nd). The *the+na* construction, with no obvious agreement on *the*, is also used to express future. What both *thelei*+V and *the na* constructions have in common is the presence of referential agreement on the second (main) verb. Whether *the* is an auxiliary or a particle at this stage is not so easy to tell. It is possible that it still bears a +V feature, but given its scopal properties it raises to a higher position. The next step of grammaticalization, then, involves a change from Move to Merge, whereby *the*, or more precisely the new form *tha*, is merged directly in a higher functional head. We take this head to be above T.

The relevant structures are as in (46) below:

(46) a. $[_M$ thelei $[_T$ grafo $[_v$ t_{grafo} . . .]]]
b. $[_M$ the $[_T$ $[_v$ $[_C$ na $+V_{lexical}]]]$ > $[_M$ tha $[_T$ $[V_{lexical}]]]$

Example (46a) is the structure without *na*, which is monoclausal. Given that *thelei* is merged in a higher functional head, the lexical verb is allowed to raise to T. This change is important as we expect that upon the development of *tha* the lexical verb can inflect for past tense. Indeed this is the case, as shown in (26a–b) above. The prediction we make is that these constructions develop much later than the 'future' *tha*. As Pappas (2001) shows, this is in fact the case. The structure in (46b), on the other hand, has *the* in M and a series of unrealized functional heads, possibly lacking a lexical V altogether. In the absence of any tense or agreement marking on *the*, the structure in (46b) is reanalysed as a monoclausal one by the language acquirer. The question is how the *the na* form which survives in some MG dialects is to be analysed. We tentatively suggest that in these dialects *the na* has been reanalysed to a single lexical item (cf. *the na* > *enna* in Cypriot Greek. See also Jarad (1997) for a similar analysis of *for to* in the history of English.).

The next question is whether *tha* can be treated as an affix, or whether we can predict that it will reduce to an affix. In our discussion of the Romance future in section 2.2, we argued that the reduction of *habere* to an affix depended on a 'head-final' structure, that is, a structure with V raising over the auxiliary

habere. If our analysis is correct, then in order for *tha* to become a suffix, we would expect systematic V fronting to a higher C position. However, this is not attested in MG. In fact, *tha* and V movement seem to be in complementary distribution (cf. Roussou 2000). On the other hand, one might wonder whether *tha* could be analysed as a prefix. If that were the case, then we would expect that *tha* is always attached to the verb, contrary to fact, as it is possible to find clitics between *tha* and V. The systematic presence of clitics blocks the adjacency required between the two heads that host *tha* and the verb. Thus reanalysis to a prefix is not possible at this stage (see also the role of clitics for the (non)reduction of *habeo* to an affix in section 2.2).

2.4 Conclusion

To conclude, in the present chapter we have considered the development of English modals, the Romance and the Greek future. The first is a case of lexical verb > auxiliary reanalysis, whereby a lexical verb becomes an auxiliary and is merged directly in the relevant functional head. The second case is slightly more complex as it also involves reanalysis of an auxiliary to an affix. We showed that this step of diachronic change is attributed to the 'head-final' nature of Late Latin grammar. The third case involves the reanalysis of an auxiliary to a particle, which is realized higher up in the clause structure. As we will show in Chapter 3, this higher position can be identified with a C head (if this is correct then this is an instance of change from V to being in the I and then in the C system). This supports our contention that grammaticalization is reanalysis 'upwards' along the functional structure. Since movement is always local and upward, categorial reanalysis is also local and upward. We also notice that this kind of change affects a small, unproductive, morphological subclass of lexical class L, which can be reanalysed as a functional class. In particular, all reanalysed verbs discussed in this chapter have common characteristics, that is, they are stative, intensional, and arguably more prone to a non-thematic interpretation as modal functional heads.

We have also argued that grammaticalization-type changes follow a 'path' (*pace* Lightfoot 1998). This 'path' is structurally defined, broadly following the Cinque (1999) hierarchy of functional categories. Moreover, the path is traversed by the loss of steps of head movement, leading to changes from Move to Merge. As we will discuss in Chapter 5, section 5.2.2, the loss of movement of L (lexical) to F (functional) can yield two possible outcomes. The first case can apply to all Ls, giving word-order changes (loss of V2, loss of V-to-I, VSO > SVO, OV > VO, etc.). It is 'downward', fully productive, and involves no clear

semantic or phonological change (that is to the L-roots). The second case can apply just to a morphologically defined subclass of L, recategorizing it as F (ME modals, *bisogna, habere, thelo*, etc.). This loss of movement is 'upward', unproductive (but sensitive to morphological sub-regularities), and associated with semantic and phonological change (the former at least directly explicable in terms of category change). Loss of movement in general is a mechanism of change due to properties of language acquirers (cf. Clark & Roberts 1993), who aim at least-marked settings. As we argued in Chapter 1, the changes from $F^*_{Move} > F_{Merge}$ and from $F^* > F$ do yield less marked parametric settings and are thus to be preferred. We will discuss this in detail in Chapter 5.

3 *C elements*

3.0 Introduction

In Chapter 2 we discussed three well-known cases which involve reanalysis of a verb to an auxiliary element, an affix, or a particle. All three cases share reanalysis of V to a T element. In the Romance case, though, *habeo* further reduced to a suffix, while in Greek *thelo* became a particle arguably in the C system, thus following the V > T > C reanalysis path. In this chapter, we turn to the grammaticalization of C elements. In the first three sections (3.1–3.3) we will consider the development of the subjunctive particle *na* in Greek, of Southern Italian *mu* and of the infinitival marker *to* in English. In section 3.4, we look at the accounts of the development of *that*-complementizers in Germanic (cf. Ferraresi 1991, 1997, Kiparsky 1995, Longobardi 1991) and in connection to this we also briefly discuss the Greek complementizer *pou*. Finally, in section 3.5 we consider the development of complementizers out of lexical verbs, and in particular out of a serial verb construction. Our analysis heavily relies on the data discussed in Klamer (2000). In this case we also show that lexical to functional reanalysis is upwards.

Section 3.1 starts with the particle *na* in Greek, as its discussion is crucial for the analysis of the elements *mu* and *to*. The development of *na* is seen in the light of the changes that took place in the history of Greek and were discussed in the previous chapter in relation to *tha*. There we showed that in diachronic terms *tha* is derived from *thelo* + *na*. Synchronically, *tha* and *na* are in complementary distribution and share a number of properties; furthermore, *na* is also in complementary distribution with the complementizers *oti* (that) and *an* (if). The Southern Italian particle *mu* is, on the other hand, not in complementary distribution with other complementizers (in particular *chi*, the 'that' complementizer). In all other respects, however, it is very similar to *na*, as we will see in section 3.2. There we will describe the development of *mu* from the Latin adverb *modo* ('in this way') and the complementizer *ut*. The development of English *to* cannot be seen independently of that of modals, as it seems that both