1

Introduction

Over the course of the last three decades, a wealth of data has been published on the origin and development of grammatical forms. The main purpose of the present work is to make this wealth accessible to a wider readership. To this end, over 400 processes relating to the evolution of grammatical categories are discussed, using data from roughly 500 different languages. (See Appendix 3 for a list of languages figuring in this book.)

The readership we have in mind for this book includes first of all linguists. Grammaticalization theory, which is the framework adopted here (see §1.1), is concerned with language use across space and time; hence the findings presented may be of help for diachronic reconstruction, especially in areas where other tools available to the historical linguist, such as the comparative method and internal reconstruction, do not yield appropriate results. The descriptive linguist will find information, for example, on how and why different grammatical meanings can be related to one another in a principled way (i.e., on how to deal with issues like polysemy and heterosemy), on why there are some regular correspondences between grammatical forms and the meanings expressed by them, or on why certain linguistic forms have simultaneously lexical and grammatical functions. Anthropologists, sociologists, and psychologists may discover that the kind of human behavior held responsible for the evolution of grammatical forms is not all that different from the kind of behavior they observe in their own fields of study.

What distinguishes this work from relevant monographs on grammaticalization theory (e.g., Lehmann 1982; Heine and Reh 1984; Heine, Claudi, and Hünnemeyer 1991; Traugott and Heine 1991a, 1991b; Hopper and Traugott 1993; Bybee, Perkins, and Pagliuca 1994; Pagliuca 1994; Heine 1997b; Ramat and Hopper 1998) is its conception as a reference work. Accordingly, an attempt was made to collect many data from as many different languages as possible and to avoid theoretical biases – as far as this is possible and feasible.

1.1 Grammaticalization Theory

Grammaticalization is defined as the development from lexical to grammatical forms[1] and from grammatical to even more grammatical forms. Since the development of grammatical forms is not independent of the constructions to which they belong, the study of grammaticalization is also concerned with constructions and with even larger discourse segments.

In accordance with this definition, grammaticalization theory is concerned with the genesis and development of grammatical forms. Its primary goal is to describe how grammatical forms and constructions arise and develop through space and time, and to explain why they are structured the way they are.[2] Technically, grammaticalization involves four main interrelated mechanisms.

(a) desemanticization (or "semantic bleaching") – loss in meaning content,
(b) extension (or context generalization) – use in new contexts,
(c) decategorialization – loss in morphosyntactic properties characteristic of lexical or other less grammaticalized forms, and
(d) erosion (or "phonetic reduction") – loss in phonetic substance.

While three of these mechanisms involve a loss in properties, there are also gains. In the same way that linguistic items undergoing grammaticalization lose in semantic, morphosyntactic, and phonetic substance, they also gain in properties characteristic of their uses in new contexts. Grammaticalization requires specific contexts to take place, and it can be, and has been, described as a product of context-induced reinterpretation. Accordingly, context is a crucial factor in shaping the structure of grammatical forms – to the extent that they may express meanings that cannot immediately be derived from their respective source forms.

It has been argued that grammaticalization is not a distinct process, since the four mechanisms can be observed to be at work also in other kinds of linguistic change (Newmeyer 1998: 248ff.).[3] There are a couple of reasons why we think that such a position is not justified. First, the main task of grammaticalization theory is to explain why grammatical forms and constructions are structured the way they are, and these four

[1] The term "grammatical forms," or "grams," roughly corresponds to what is also referred to as "functional categories."

[2] Newmeyer (1998: 240) raises doubts about whether we are really dealing with a theory here, and he rightly observes that much of the relevant literature on this subject is not very helpful on deciding this issue.

[3] Newmeyer (1998: 260) refers to desemanticization as "appropriate semantic change," to decategorialization as "downgrading analysis," and to erosion as "phonetic reduction."

mechanisms, as opposed to many other conceivable mechanisms, have been found to be relevant to achieve such explanations. Thus, irrespective of how one wishes to define a "distinct process," one is led to conclude that these mechanisms are part of one and the same explanatory framework.

Second, grammaticalization, as conceived here, is above all a semantic process. This process is context dependent, and grammaticalization can therefore be described in terms of context-induced reinterpretation. Not every reinterpretation leads to the rise of grammatical meanings. Rather, it is only when forms for concrete (e.g., lexical) meanings are used to also express more abstract (grammatical) meanings that grammatical forms emerge; for example, when a form used for a visible object (e.g., the body part 'back') is used also to refer to a nonvisible item (the spatial notion 'behind'), or a form used for an action ('go to') is used also to refer to a grammatical notion (future tense). On account of its specific directionality, context-induced reinterpretation has been described in terms of metaphorical transfer, leading, for example, from the domain of concrete objects to that of space, from space to time, from ("real-world") space to discourse space, and so on.

Desemanticization thus results from the use of forms for concrete meanings that are reinterpreted in specific contexts as more abstract, grammatical meanings. Having acquired grammatical meanings, these forms tend to become increasingly divergent from their old uses: they lose in categorial properties characteristic of their old uses, hence undergoing decategorialization, and they tend to be used more frequently, to become more predictable in their occurrence, and, consequently, to lose in phonetic substance. Thus, the four mechanisms are not independent of one another; rather, desemanticization precedes and is immediately responsible for decategorialization and erosion. There are a few cases where it has not yet been possible to establish that decategorialization really followed desemanticization in time, and we do not wish to exclude the possibility that in such cases the two may have occurred simultaneously. However, such cases appear to be exceptional: new grammatical meanings arise, and it usually takes quite some time before any corresponding morphological, syntactic, and/or phonetic changes can be observed. In many languages, prepositions unambiguously serving a grammatical function still have the morphosyntactic structure of their earlier uses as adverbial phrases (cf. English *by means of*, *in front of, with respect to*) or verbal phrases (cf. Chinese *ZAI* '(to be) at'; Alain Peyraube, personal communication), and tense or aspect auxiliaries may still behave morphosyntactically largely like lexical verbs even if they have lost their lexical semantics and serve exclusively as functional categories (cf. English *be*

going to, used to, keep (doing), etc.). To conclude, there is evidence to suggest that grammaticalization can be defined as a distinct process.

It is sometimes assumed that grammaticalization invariably involves lexical categories; that is, that it is confined to the development from lexical to grammatical forms. This view tends to ignore that such cases account for only part of what falls under the rubric of grammaticalization. Equally commonly, as we will see in the course of this work, items already part of the inventory of grammatical forms give rise to more strongly grammaticalized items. Prepositions often develop into conjunctions, temporal conjunctions tend to give rise to causal or concessive conjunctions, demonstrative determiners develop into definite articles or relative clause markers, verbal perfect inflections may become past tense markers, and so forth – all developments that take place within the domain of functional categories. Such developments are distinguished mainly from developments involving lexical categories by the difficulty of identifying and reconstructing them.

Grammaticalization is a unidirectional process; that is, it leads from less grammatical to more grammatical forms and constructions. However, this process is not without exceptions: a number of examples contradicting the unidirectionality principle have been found (see, e.g., Joseph and Janda 1988; Campbell 1991; Ramat 1992; Frajzyngier 1996; and especially Newmeyer 1998: 260ff.). Yet, as acknowledged by most of the scholars who have identified exceptional cases, such examples are few compared to the large number of cases that conform to the principle[4] (cf. Haspelmath 1999, 2000: 249). Furthermore, they can frequently be accounted for with reference to alternative forces,[5] and finally, no instances of "complete reversals of grammaticalization" have been discovered so far (cf. Newmeyer 1998: 263).

Grammaticalization begins with concrete, lexical forms and constructions and ideally ends in zero – that is, grammatical forms increasingly

[4] Cf., e.g., Harris and Campbell (1995: 338), who summarize this situation thus: "there is a strong tendency for grammaticalization to proceed in one direction, though it is not strictly unidirectional." Similarly, Joseph and Janda (1988: 198–200) observe that cases of demorphologization, a process that would contradict the unidirectionality principle, are rare and not seldom controversial. Finally, Newmeyer (1998: 275–6, 278) observes that cases conforming to the unidirectionality principle ("downgradings") "have occurred at least ten times as often as upgradings," and he concludes, "I suspect that, for whatever reason, there is a *general* directionality to the semantic changes observed in grammaticalization" (emphasis in original).

[5] Such forces may be morphophonological or morphosyntactic in nature, but they may as well relate to specific sociocultural factors. Burridge (1995) discusses an example of reversed directionality in Pennsylvania German, where a modal auxiliary developed into a lexical verb, *wotte* 'wish'. As Burridge shows, one factor contributing to this development can be found in the particular Mennonite religious principles held by the speakers of Pennsylvania German.

lose in semantic and phonetic content and, in the end, they may be replaced by new forms; grammaticalization has therefore been described as a cyclical process (Givón 1979a; Heine and Reh 1984).[6] While there is some evidence to support this assumption, we have to be aware that, first, a grammaticalization process can stop at any point of development and, second, "worn-out" grammatical forms are not necessarily replaced by new forms. Thus, the metaphor of a grammatical cycle, though useful in certain cases, should not be generalized since it often does not apply for some reason or other.

In a number of works, grammaticalization is described as a process that involves the reanalysis of grammatical categories.[7] Other authors have argued that there is no necessary relationship between grammaticalization and reanalysis (see especially Haspelmath 1998). In fact, reanalysis has been defined in a number of different ways (cf. Langacker 1977; Heine and Reh 1984; Harris and Campbell 1995: 61–96; Haspelmath 1998; Newmeyer 1998: 241–51). Whether grammaticalization involves reanalysis has turned out to be essentially a theory-dependent issue. To avoid any further confusion on this issue, we prefer to exclude "reanalysis" from our terminology of grammaticalization theory.

1.2 Problems

Grammaticalization is a complex subject matter; it relates in much the same way to diachronic and synchronic linguistics as to semantics, syntax, and morphology, and it is rooted in cognition and pragmatics. Obviously, an endeavor such as that found here is an ambitious one – one that has to take care of a wide range of problems. In this section we deal with the most serious of these problems in turn.

The findings presented in this work are meant to highlight processes of human behavior that can be observed across cultures; yet, these findings are based on data from hardly more than one-tenth of the world's languages. One may therefore wonder what justification there is to call this work a "world lexicon." Our main reason is this: underlying human behavior there appears to be a strategy of linguistic processing whereby more abstract functions are expressed in terms of forms for concrete concepts. We expect, for example, that in some unknown language there are

[6] Givón (1979a: 209) proposed the unidirectional cycle in (i), where the end point (Zero) marks the beginning of a new cycle again leading from Discourse to Zero:
(i) Discourse > Syntax > Morphology > Morphophonemics > Zero.

[7] Newmeyer (1998: 238), for example, argues, "The standard definition of grammaticalization incorporates the notion of reanalysis; no definition that does not do so seems particularly useful."

ways of expressing temporal concepts in terms of spatial ones, spatial relations in terms of forms for concrete concepts (such as body parts or salient landmarks), aspectual contours of events in terms of forms for actions and motions, or functions concerning the organization of texts in terms of linguistic forms for spatial or temporal deixis. Languages differ considerably in the way and the extent to which this strategy has given rise to grammaticalized constructions; nevertheless, we expect the effects of this strategy to be essentially the same across languages, including languages that are still undocumented.

Throughout this work we are concerned with the relation between two kinds of concepts, which we refer to as the "source" and "target" entities of grammaticalization. We convey the impression in this account that there is always a unidirectional development leading from one distinct entity to another entity. But this is not only a simplified account; it is also at variance with much of what we have argued for elsewhere, namely that, rather than being a development in discrete steps, grammaticalization must be described as a continuous or, more precisely, as a chainlike development (Heine 1992). To achieve the goal of having a treatment of grammaticalization processes in the form of a lexicon, we were forced to reduce continuous, chainlike structures to two salient uses of forms, viz., source and target uses.

Most of the over 400 grammaticalization processes discussed in this book are based on fairly reliable reconstruction work, but in some cases the evidence available is not yet satisfactory. We have pointed out such cases under the relevant entry.

A number of developments leading to the evolution of grammatical categories do not involve linguistic units like words or morphemes (Heine 1993; Bybee et al. 1994; Bisang 1998a); rather, they concern more complex conceptual entities, such as phrases, whole propositions, or even larger constructions. For example, the temporal conjunction *taátenu* 'then' of Kxoe, a Central Khoisan language of Namibia, is historically a clause meaning 'when it is like that' (see (1)).

(1) *ta- á- te- nu xaváná ǁé kúùn-à- tè . . . .*
be:thus-JUNC-PRES-when again 1:M:PL go- JUNC-PRES
'Then we went again. . . .'

A much better known example concerns the evolution of aspect and tense categories, where two or more different linguistic forms may simultaneously be involved: an auxiliary (e.g., *be* or *have*), a nonfinite marker (e.g., an infinitival, participial, or gerundival marker), and perhaps also a locative marker. Tense and aspect constructions in a number of languages worldwide not uncommonly involve three distinct morphological

elements, the English future marker *be going to* being a paradigm example. Another European example is the Latin verb *habere* 'to have', which in the Romance languages has given rise to perfect markers on the one hand and to future markers on the other. What accounts for this divergent development? The verb *habere* was not itself grammaticalized; rather grammaticalization involved entire periphrastic constructions, or event schemata: the construction *habere* + perfect passive participle gave rise to perfect expressions, while *habere* + infinitive periphrasis was responsible for the development of future constructions. In a lexicon project like the present one, such propositional structures had to be reduced to the salient segments of the constructions concerned, such as the *habere*-markers figuring in the expression of future tenses in Romance languages.

A related problem that we encountered concerns what one may call "complex grammaticalization": a more complex linguistic structure can assume a grammatical function without involving the grammaticalization of any particular item figuring in this structure. Take (1) again: which of the various items figuring in the Kxoe word *taátenu* should be held responsible for the relevant grammaticalization? The most obvious answer would be that, rather than any particular item, the structure as a whole is responsible. In a treatment of the kind attempted here, however, which rests on the assumption that there is essentially a one-to-one correspondence between source and target, such an answer is not entirely satisfactory. What exactly should the lexicon entry be that takes care of this grammaticalization? Or take the following example: one widespread way of developing expressions for the grammatical concept of a comparative of inequality is to juxtapose two propositions that are in a polar contrast – one expresses the standard of comparison and the other the comparative notion. This opposition may be either antonymic, as in (2), or marked by the distinction of positive versus negative, as in (3).

Cayapo (Stassen 1985: 184)

(2)	*Gan*	*ga*	*prik,*	*bubanne*	*ba*	*i*	*pri.*
	you	you	big	but	I	I	small

'You are bigger than I am.'

Abipon (Stassen 1985: 184)

(3)	*Negetink*	*chik*	*naâ,*	*oagan*	*nihirenak*	*la*	*naâ.*
	dog	not	bad	yet	tiger	already	bad

'A tiger is more ferocious (lit.: 'bad') than a dog.'

What is grammaticalized in such constructions is not a specific element but rather some propositional relation, viz., *be big* versus *be small*, or *be*

bad versus *not be bad*. In a treatment like this book, which is concerned with segmentable linguistic forms, functions expressed by means of pragmatic or syntactic relations between forms without involving morphological segments of necessity had to be excluded.

The sentence in (3) raises another question: At which point can we say that grammaticalization has been concluded? Can we really say that (2) and (3) are suggestive of a completed process of grammaticalization, or do they merely represent contextually induced interpretations that are irrelevant for the grammatical structures of the languages concerned? A number of tests have been proposed in grammaticalization theory to deal with this question; frequently, however, the information available on a given language is not sufficient to allow for a successful application of these tests. In such cases we have decided to adopt the solution proposed by the author(s) dealing with that language.

In some cases we decided to rely on comparative findings to determine whether a grammaticalization process has been concluded. For example, one of our entries has the form ONE > INDEFINITE, according to which the cardinal numeral for 'one' may grammaticalize to indefinite articles. Now, it has been argued, for languages like English (*a*(*n*)) or German (*ein*), for example, that the two, numeral and indefinite article, are the same, their difference being due to contextual or other factors; that is, that the relevant entry is not an instance of grammaticalization. That the two meanings are in fact different is suggested by comparative observations. Thus, there are languages where a given linguistic item serves as an indefinite marker but not as a numeral, and, conversely, there are many languages where a given item denotes the numeral 'one' but not indefinite reference. We take such observations as evidence that ONE and INDEFINITE are in fact different concepts, even if in some languages the same or a similar word is used for both.

Another problem concerns the directionality of grammaticalization and how to achieve historical reconstruction. How do we know that INDEFINITE is historically derived from ONE rather than the other way around? In this case, there is diachronic evidence to give an answer: in some languages, including a number of European ones, there is a marker that is used for both the numeral 'one' and the indefinite article, and by using historical records it is possible to establish that at some earlier stage in the development of these languages the item only served as the numeral expression before its use was extended to also designate indefinite reference. Now, since grammaticalization is essentially unidirectional, we are led to assume that in languages where no historical records are available the evolution was the same.

Even in the absence of historical documents it is possible to reconstruct directionality of change by using the mechanisms sketched in the preceding section. For example, decategorialization has the effect that the element concerned loses in morphosyntactic properties characteristic of its less grammaticalized (e.g., lexical) source, such as the ability to take modifiers or inflections, and it shifts from a category having many members (e.g., an open class) to a category having only few members (a closed class). Erosion again means that that element tends to become shorter and/or phonetically less complex, to lose the ability to receive distinct stress or tone, and so on. Thus, if we find two different uses of a given element, or two etymologically related elements, where one shows the effects of decategorialization and erosion whereas the other does not, then we can argue that the latter is less grammaticalized and then reconstruct a directionality from the latter to the former, rather than the other way around. Even if we had no previous knowledge of the history of English we could nonetheless establish that the indefinite article *a(n)* is a later development form of the numeral *one*, rather than the reverse, since the article exhibits a number of effects of decategorialization and erosion while the numeral does not. In this text we use this kind of evidence for reconstruction in addition to any kind of historical evidence that may be available.

Grammaticalization does not occur in a vacuum, and other forces also shape the evolution of grammatical forms, language contact being one. The rise of a new grammatical expression may be the result of grammaticalization, but it may also be due to the influence of another language. The question of whether, or to what extent, a given development is from language-internal as opposed to language-external factors can frequently not be answered satisfactorily. Recent studies suggest that both are often simultaneously involved.

These observations led us to the question of whether any restriction in the kind of linguistic transmission should be imposed when selecting the data to present in this volume. For example, should instances of grammaticalization that clearly occurred due to borrowing be excluded? Should we separate such cases from instances of grammaticalization that have to do with continuous transmission within a given language?

A perhaps related issue concerns pidgins and creoles, which are a gold mine for students of grammaticalization, and throughout the 1990s a wealth of publications appeared demonstrating the relevance of grammaticalization theory to the study of these languages (see especially Baker and Syea 1996). With the rise of pidgins and creoles, the question again arises as to whether we are dealing with "natural" forms of transmission

and, if yes, whether grammaticalization processes behave the same way whether they have taken place, for example, between earlier and later forms of British English or between British English and Krio CE or Tok Pisin PE. The policy adopted here is to take all these kinds of data into account, at least as far as they are in accordance with principles of grammaticalization observed in "natural" language transmission. More recent research suggests that grammaticalization in pidgins and creoles does not behave essentially differently from that found in other languages. The reader is in a position to identify instances of borrowing or pidginization, or creolization, on the basis of the exemplification provided in this book.[8]

The terminology used to refer to grammatical categories differs from one author to another and from one language to another. Although we have tried to standardize terms, in many cases, this turned out to be impossible because of insufficient information. It is therefore to be expected that, in accordance with the conventions adopted by the relevant authors, one and the same grammatical function may be referred to by entirely different labels, both within a given language and across languages.

The quality of the data provided in this work crucially depends on the kind of information contained in the published sources that we were able to consult. Frequently it turned out that the information was not satisfactory. For example, when dealing with a verb as the source for a certain grammatical category, it is not enough to consider the lexical semantics of that verb; which grammaticalization it undergoes may depend entirely on its valency. In Southern Sotho, a Bantu language of Lesotho and South Africa, we find, among others, instances of grammaticalization like those presented in (4).

Southern Sotho (Bantu, Niger-Congo; Doke and Mofokeng [1957] 1985)

(4)

Verbal source		Grammatical form	
-ea	'go (to)'	*-ea-*	immediate future tense
-tla	'come (to)'	*-tla-*	future tense
-tsoa	'come from'	*-tsoa-*	immediate past tense

These examples suggest that it is not the deictic semantics of 'come' or 'go' that can be held responsible for the particular functions the result-

[8] Pidgin (P) and creole (C) examples are marked by adding abbreviated labels after the language name. For example, "CE" stands for "English-based creole" (see Abbreviations). Note that the classification underlying this usage is a crude one, since terms like "English-based," "Portuguese-based," etc. are not unproblematic, and the boundary between pidgins and creole languages is not seldom fuzzy.