Part I

Theoretical Prerequisites

1 Introduction

1.1 The Topic of This Book

Languages differ in a vast number of features,[1] but linguistic diversity is unevenly distributed in space. This is the insight behind *areal typology*, the linguistic discipline that seeks to discover areal patterns in the distribution of linguistic features in the world's languages. Some asymmetries in the distribution of particular linguistic features are easily discovered, especially on small-scale levels. This can be done when an area of long-established language contact in which a set of features is present is contrasted with its surroundings in which that set is absent. For example, it has long been known that languages of Mainland Southeast Asia share a number of features, irrespective of the language family they belong to: Some are Sino-Tibetan (e.g. Burmese and Cantonese), some belong to the Austro-Asiatic family (e.g. Vietnamese and Khmer), and others to Tai-Kadai (e.g. Thai and Lao). The shared features include pitch accent, isolating structure, the presence of honorifics and the SVO order of syntactic constituents, and this clustering of features is explained by well-documented and long-lasting contacts between speakers of languages in that area. In a similar manner, we know that languages of the Balkans share a number of characteristics, including the lack of infinitive, the presence of the vowel "schwa" (ə), the syncretism of the genitive and dative cases, the postposed definite article and the analytical future tense construction (usually formed with the auxiliary verb meaning "to want" and the present (subjunctive) of the lexical verb). Historical grammars of individual Balkan languages (Albanian, Romanian, Greek, Bulgarian and Macedonian) explain how and even when such features spread in the area and show that they cannot be inherited from the common proto-language (Proto-Indo-European). All of this is the subject of *contact linguistics*, which seeks to explain how languages borrow

[1] In this book, we will use the term *feature* to refer to any characteristic of languages that, for whatever reason, a typologist may find relevant. The term is also used to refer to a special subset of morphosyntactic features, such as number, gender, case, singular, dual, masculine, feminine, etc. (Corbett 2012). For these features we will also use the terms *categories* and *subcategories*.

structural features in situations of intensive contact (see, e.g., Thomason and Kaufman 1988). However, we can also look at the distribution of individual linguistic features (or of clusters of features) from the global point of view. For example, we can ask how languages with pitch accent (or without infinitives, or with "schwa"…) are distributed over large areas where no historical contacts are documented. We can also try to determine whether the presence or absence of certain features in particular areas are correlated, and whether the presence (or absence) of individual features in an area predicts (or implies) the presence or absence of other features in that area. Of particular interest is determining whether the areal distribution of particular linguistic features is correlated with the distribution of languages belonging to certain language families, i.e., whether genetic factors play a role in the explanation of the attested distribution of features. Such investigations are the proper domain of areal typology.[2]

This book is about the areal distribution of various patterns of *agreement*, which is a type of grammatical rule that some languages have, while others lack it, and those languages that have it may have it in different domains and for different grammatical categories.[3] In the following chapters, we shall mostly limit ourselves to two syntactic domains of agreement: (1) the noun phrase (NP), in which modifiers agree with the head noun; for example, in English there is number agreement shown in the opposition between the NPs *this book* and *these books*; and (2) the clause, in which the verb usually agrees with one or more of its arguments; for example, in English the verb agrees with its subject in number and person, as shown in the opposition between *The man sings* and *We sing*. In the NP domain, we shall systematically investigate the areal distribution of agreement patterns for the categories of gender, number, case and (in possessive NPs) person. In the domain of the clause, we shall be looking at the areal distribution of agreement patterns for the categories of person, number and gender. Our conclusions will be drawn from a systematic investigation of a sample of 300 languages, constructed so as to be representative both genetically (by including languages from many different families) and areally (by including languages from many different parts of the world). However, examples will occasionally be presented from languages not in that sample, since some typologically relevant data were collected also from languages that could not be included in the sample in order not to make it areally or genetically biased.

[2] As Balthasar Bickel, who considers this type of enterprise as "twenty-first-century typology," succintly put it (2007: 239): "Instead of asking 'what's possible?', more and more typologists ask 'what's where why?'. Asking 'what's where' targets universal preferences as much as geographical or genealogical skewings, and results in probabilistic theories stated over properly sampled distributions."

[3] For a definition of agreement used in this book see Chapter 2.

To the best of our knowledge, this is the first monograph devoted exclusively to this topic, which is not to say that it is the first attempt to answer the questions which it raises. Global asymmetries in the distribution of particular categories and patterns associated with agreement have long been noted in linguistic literature. Since the inception of linguistic typology as a discipline it has been known that isolating languages of Mainland Southeast Asia and parts of Western Africa have uninflected verbs and, consequently, no verbal agreement.[4] Gerlach Royen, in his 1030-page monograph on the nominal classification systems in the languages of the world (Royen 1929) noted that gender, as a category, has a very uneven distribution in the world's languages, that it is quite common in some parts of the world while being absent (or nearly so) in others. Unfortunately, Royen's work was largely neglected,[5] but after the Second World War, with the development of contact linguistics and the notion of language areas or *Sprachbünde*, patterns in the worldwide areal distribution of certain agreement features were noted, especially with respect to gender and person agreement; studies on large language areas include Emenau (1956) for the Indian Subcontinent, Campbell, Kaufman and Smith-Stark (1986) for Mesoamerica, and Sherzer (1973) for a number of areas in North America (especially for the languages of the Northwest Coast). In these works the presence or absence of specific agreement patterns plays a role in the definition of individual language areas.

By the late 1980s, most large linguistic areas had become recognized, as well as the global skewings in the distribution of many typologically prominent linguistic features (Dryer 1989). "It has become clear… that hardly any typological variable, and only some combinations thereof, is evenly distributed in the world" (Bickel 2007: 243). The time was ripe for a new synthesis of areal typology, and this came in the form of Johanna Nichols' award-winning book *Linguistic Diversity in Space and Time* (Nichols 1992). In that book, Nichols examined the distribution of many linguistic features among language areas and linguistic families, with the specific purpose of determining their areal and genetic stability. Among the surveyed features were, e.g., the presence of inclusive/exclusive opposition in pronouns, word order, the alienability contrast in possessive constructions, but also some grammatical categories, including

[4] Wilhelm von Humboldt, in a talk given to the Prussian Academy of Sciences in 1827 (reprinted in von Humboldt 1963), had already noted that the dual (as a subcategory of the category of number) has a limited distribution in the world's languages. Moreover, he was aware that in some languages (especially in the languages of Polynesia and the Philippines), the dual is expressed only in pronouns, in others it is found only in nouns (e.g. in Totonac), while only in some languages and areas it characterizes several parts of speech, thus participating in agreement (in this group he rightly included Indo-European and Semitic languages as well as Greenlandic).

[5] For a general history of nominal classification in languages of the world, see Kilarski 2013.

number and gender. Number oppositions and their neutralizations were examined without considering whether languages exhibit number agreement or not, but the distribution of gender agreement was very carefully investigated, and it was concluded that gender is one of the genetically most stable categories in language. That is, gender characterizes whole language families, and languages that belong to gendered families are unlikely to lose gender agreement, while those that belong to genderless families are unlikely to acquire it. This insight is, as we shall see further below, very important for our more general investigation into the areal distribution of agreement systems, and Nichols' findings are fully confirmed by our data.

The areal distribution of some of the features discussed by Nichols, although those features are not involved in agreement themselves, has a direct bearing on the areal distribution of agreement systems. For example, Nichols has shown that head-marking languages (those in which the head of a syntactic construction is morphologically marked rather than the dependent element) tend to cluster in areas she calls "hotbeds," and that they are unexpectedly more common in the New World (Australia, New Guinea, Oceania and the Americas) than in the Old World (Africa and Eurasia). On the clause level, head-marking languages do not have case, but rather indicate grammatical relations by cross-referencing on the verb. This means that case agreement cannot be found in head-marking languages because, trivially, there cannot be case agreement where there is no case. In a similar manner, modifiers of the noun (adjectives, demonstratives, numerals and articles) are dependent elements of the NP, while the noun is its head. In a head-marking language the modifiers of the noun will not agree with it in categories such as number and gender (although verbs may agree with one or more arguments on the clause level), so languages with extensive adnominal agreement will typically not be head-marking languages, and the areal distributions of the two types will greatly differ. Therefore, although agreement as a grammatical phenomenon was not in the focus of Nichols' investigation, her work contains many insights into the areal typology of agreement systems.

Considering that many patterns in the areal distribution of agreement systems were already noted in the literature, the question arises whether a monograph about areal typology of agreement is really necessary, or even useful. Why should one read a book on the areal typology of agreement written by a general typologist, when one can read about patterns of agreement in works written by specialists on various languages and families spoken in different parts of the world? Indeed, information about the distribution of individual categories involved in agreement, such as number, gender, person and (less often) case, can be found in several specialized monographs dealing with large, often continent-sized macro-areas. We can mention Adelaar (2004) for languages of the Andes (and some neighboring

areas), Dixon and Aikhenvald (1999) for languages of the Amazon, Dixon (2002) for Australian languages, Crowley (1998) for languages of the Pacific, Welmers (1973), Gregersen (1977) and Heine and Nurse (2000) for African languages, Van Driem (2001) for languages of the Himalayas, Foley (1986) for Papuan languages, Mithun (1999) for languages of North America and Suárez (1983) for Mesoamerican languages. The series of monographs on the "languages of the Soviet Union" (e.g. Vinogradov 1967) is still very useful for the languages of Siberia and the Caucasus, and for the latter there are also monographs by Klimov (1986) and Hewitt (2004). Comrie (1981) covers all of the languages of the former Soviet Union. However, data from such sources is often difficult to use, since the categories involved in agreement are not defined consistently by different authors. For example, in some sources the term "gender" is reserved for languages in which nominal classification is based on sex-based oppositions, while in others it refers to any system of nominal classification which manifests itself in agreement. Likewise, in some sources (especially in the Africanist tradition, see Creissels 2006), morphemes that would otherwise be called person agreement markers are treated as independent pronouns, which makes it difficult to establish the distribution of languages with verbal person agreement using the same standards in all languages. To be able to tell whether a particular agreement pattern is common in a given area, one first has to make sure that the source one is consulting for the languages of that area uses the same definition of the relevant agreement pattern, and uses it consistently. However, that is very often not the case.

Surveys of areal distribution of individual agreement patterns are rare in the existing literature. Nevertheless, Hurskainen (2000) gives a useful overview of gender systems in African languages, and for gender systems in Eurasia there is a paper by Juha Janhunen (2000) and a chapter of my book on gender in Indo-European (Matasović 2004: 191–211). Michael Rießler's PhD thesis (2011, published in 2016) showed areal patterns in adjective attribution in the languages of Northern Eurasia. It basically confirmed that agreement in the NP is rare in that area and virtually limited to Indo-European languages. Balthasar Bickel and Johanna Nichols (2009) looked at the global distribution of various aspects of case systems, and one of the features they surveyed was case agreement, which they call "case spreading" (i.e. the spreading of case marking from the head noun to other elements of the NP). They did not find any areal biases in the distribution of languages with case agreement, "perhaps because the datasets for these variables are so small" (Bickel and Nichols 2009: 489; they examined only 63 languages). To the best of our knowledge, there are no areal–typological studies about the distribution of languages with number agreement in either the NP or the verb. This is perhaps surprising, since the last few decades have seen a growth of interest in the areal typology

of languages, and the distribution of many linguistic features has been investigated and plotted on language maps.

With the publication of the "World Atlas of Linguistic Structures" (or WALS) in 2005 areal typology received a big boost as a linguistic discipline. Some of the maps in WALS bear directly on the questions we are dealing with in this book. Greville Corbett's Map 30A (Corbett 2013a) shows the distribution of languages with gender worldwide. It demonstrates that there are areas in which this category is rare (e.g. Northern Eurasia, the Andes, Southern Australia, etc.), as well as those where it is common (e.g. Southeast Eurasia, most of sub-Saharan Africa). However, it does not show the syntactic domains in which the represented languages have gender agreement (some have it in the NP, some on the verb, and some in both domains), and the same holds for the other maps dealing with gender (Maps 31A and 32A, see Corbett 2013b and Corbett 2013c). Map 29A "Syncretism in Verbal Person/Number Marking" by Matthew Baerman and Dunstan Brown (Baerman and Brown 2013) can be used to show the areal distribution of languages without person marking (these are common in Mainland Southeast Asia, parts of sub-Saharan Africa and Southern Australia), but only languages where the subject person is marked on the verb are included in the category of languages with verbal person marking. Languages in which the verb agrees with its object, but not its subject, although they are rare, were not recorded. Moreover, that map cannot be used to find languages that have verbal agreement in gender and number, but not in person. Map 58A (Bickel and Nichols 2013) shows the distribution of languages with obligatory possessive inflection, and while this map shows an indication of the distribution of languages with person agreement in the NP, the two sets cannot be equated, since not all the languages with obligatory possessive inflection have person agreement with the possessors (see our discussion in Section 4.6).

The maps contained in WALS do show us the distribution of languages that have case marking (e.g. Map 49A "Number of Cases," Iggesen 2013), but they do not show us the distribution of languages with case agreement, and those are, obviously, a subset of languages that have the category of case. There are also maps from which the reader can gather the distribution of languages with number marking in the NP, but not the distribution of languages with number agreement, and it is impossible to get the information on the areal distribution of languages in which number is marked on the verb (separately from person/gender). On the whole, the features surveyed in WALS are strictly defined so as to assure that identical phenomena are compared across languages, but theoretical approaches to different features may and often do differ: for example, morphological marking for some categories surveyed in WALS was limited to affixes, while for others it also included clitics. Finally, it must be mentioned that the sizes of the samples of languages used for different features differ dramatically. Thus, we know more about the distribution of languages with

nominal plurality (Map 33A, "Coding of Nominal Plurality" (Dryer 2013), with 1066 languages) than about the distribution of languages in which the coding of nominal plurality is optional (Map 34A, "Occurrence of Nominal Plurality" (Haspelmath 2013b), with 291 languages). Therefore, although WALS is a major step in the development of areal typology as a discipline, the data contained in it cannot be directly used to answer the questions raised in this book.

Having seen that answers to questions we would like to ask cannot be found in the literature, it remains to be seen if those answers are interesting and deserve the effort of looking for them. This is what the rest of this book is dedicated to.

1.2 Outline of the Book

After a brief introduction to the topic of our investigation, let us present the outline of the remainder of this book by chapters.

Chapter 2 ("What Is Agreement?") starts with the general definition of agreement mostly found in textbooks and reference works ("systematic covariance between a semantic or formal property of one element and a formal property of another" [cf. Corbett 2006: 4; Wechsler 2015: 309]). After discussing this definition, we attempt to make it more operational for typological sampling, especially for determining whether a language has agreement or not, as well as for determining the typological parameters according to which languages with agreement differ.

Chapter 3 ("Domains of Agreement and Categories Involved") systematically discusses the syntactic domains in which agreement is found (chiefly the NP and the clause, although the instances of agreement within the domain of the sentence and the discourse are also briefly analyzed). It is argued that the most common agreement pattern within the domain of clause is verbal agreement (i.e. the pattern where verbs are targets of agreement), but several typologically unusual cases show that this is not the universal rule. There follows a discussion of grammatical categories involved in agreement, and we focus on the cross-linguistically most common patterns, including agreement in gender, number, person and case. Correlations between different agreement categories are also discussed, such as the universal claim that languages with gender agreement always have number agreement, and some counter-examples to such universals are presented.

Chapter 4 ("Problems with Agreement") deals with a number of phenomena that are sometimes not considered to instantiate agreement (case agreement and person agreement in possessive constructions), as well as with some constructions that are not universally accepted as agreement (constructions with omissible controllers and those with referential targets). This is important, as

it is necessary to determine in advance which types of constructions will be counted as instantiating agreement for any areal typology to make sense. This can only be achieved if we make sure that strictly defined and identical phenomena are compared across languages.

Chapter 5 ("Grammatical, Ambiguous and Anaphoric Agreement") attempts to extend Siewierska's (1999, 2004) typology of verbal agreement (in person) to adnominal agreement as well. This is important, as we wanted to make sure that verbal and adnominal agreement are truly comparable phenomena. We argue that comparing types of agreement patterns in both domains (the clause for verbal, the NP for adnominal agreement) makes cross-linguistic sense only if we limit our investigation to ambiguous agreement, and this is generally done in the rest of the book. However, since grammatical verbal agreement is a well-defined notion, and it is clearly relevant in any discussion of the typology of agreement systems, we have decided to examine its areal distribution as well.

Chapter 6 ("Marginal Agreement") discusses instances where a language has only a few lexical items that show agreement in some particular domain. It is clear that, in a typological investigation that aims to show some general patterns in the distribution of agreement features, one needs to identify languages where a particular type of agreement is marginal (e.g. adnominal number agreement in Hungarian, which is limited to a couple of lexical items), and treat them differently from those in which agreement is a pervasive phenomenon, as in Italian, where nearly all adnominal modifiers agree in gender and number with the head noun. In our study, we tried to consistently identify marginal agreement patterns in all languages included in our sample, and to test whether the statistical generalizations we found depend on the inclusion or exclusion of such languages.

Chapter 7 ("The Sample of Languages") explains in some detail the principles that guided us in the selection of languages in the 300-language sample on which the present investigation is based. Since the aim of our study was to show that certain patterns of agreement were unexpectedly rare in some macro-areas, we wanted to make sure, first, that areas were defined independently, before the sample of languages was determined, and, secondly, that the language sample was representative, in the sense that every macro-area should contain a number of languages proportional to its overall linguistic diversity.

Chapter 8 ("Areal and Genetic Patterns in Agreement Systems") looks at the areal and genetic distribution of different agreement patterns. The chapter is subdivided into sections, and each section deals with the distribution of agreement patterns in one macro-area (Eurasia, Africa and the Middle East, North America, South America, Australia and Oceania). Each section is recapitulated by a table showing which language families in a particular macro-area can be characterized as either having or lacking a particular agreement pattern. The results presented in this chapter are not based just on the analysis

of the 300 languages in our sample, but also on a comprehensive survey of the relevant typological and areal linguistic literature.

Chapter 9 ("Typological Correlations in Agreement Systems") applies statistical analyses to establish which agreement patterns are correlated, irrespective of the areas in which they are found. For example, we show that the rareness of languages with adnominal agreement without verbal agreement is statistically unexpected, and that languages with grammatical verbal agreement (i.e. where the controller of verbal agreement is obligatorily present in all clauses) regularly also have some adnominal agreement. On the other hand, we were unable to show any clear correlation between the presence or absence of agreement and word order patterns.

Chapter 10 ("Diachronic Patterns in the Development of Agreement") is, understandably, rather speculative, as the history of the majority of the world's languages remains unknown. However, we have attempted to offer a number of historical hypotheses that could explain why the geographical distribution of certain agreement patterns appears to be a priori unexpected. Since verbal agreement has been shown to be very common and evenly distributed among the world's languages, the crucial fact in need of an explanation is the distribution of languages with adnominal agreement, which is areally rather limited. We look at a number of well-documented or reasonably well-reconstructed cases, including Zande (a Ubangian language), Nilo-Saharan, Daly languages of North Australia, Proto-Indo-European, etc., and discuss the attested and probable paths in the development of adnominal agreement. It is argued that agreement often spreads from the clausal domain, where it is pragmatically motivated, to the domain of the NP, where it is largely redundant. However, adnominal agreement quite often also arises in situations of intensive language contact, e.g. in the case of Baltic Finnic languages which borrowed case agreement from neighboring Indo-European languages, or in the case of some Mande languages in West Africa, which borrowed number agreement from neighboring Gur languages. On the other hand, since all agreement patterns can be lost at any time, due to phonological erosion of agreement markers, or due to syntactic changes in a language, the rare language type with adnominal agreement and without verbal agreement is bound to develop occasionally and unpredictably.

Finally, Chapter 11 ("Conclusions") summarizes our results.