A typological database of agreement

Carole Tiberius, Dunstan Brown, Greville Corbett

Surrey Morphology Group
Linguistic and International Studies
University of Surrey
Guildford, Surrey, UK, GU2 7XH
c.tiberius,d.brown,g.corbett@surrey.ac.uk

Abstract

This paper discusses the construction of a typological database of agreement on the basis of fifteen languages taken from different language families so as to maximise diversity. For each of these languages, the database will contain detailed information about agreement controllers, targets, domains, categories, and conditions. Thus the database is designed to help us to develop a general typology of agreement systems which predicts what is, and what is not a possible agreement system in natural language. This is primarily a theoretical aim, but the database may also have practical applications in that agreement has implications for the design of parsers in natural language systems.

1. Introduction

This paper discusses the construction of a typological database of agreement. Agreement is a puzzling phenomenon found widely in languages of different types, from the familiar such as English, German, and Russian to the more exotic such as Tsakhur and Mayali. "The term agreement commonly refers to some systematic covariance between a semantic or formal property of one element and a formal property of another." Steele (1978:610). Consider the English example:

(1) The system work-s
system.3.SG work-3.SG
'The system works'

Here we have a verb agreeing with the noun phrase in person and number. Agreement in English is in some ways rather straightforward. In other languages, the phenomenon is more complex. For example, in Upper Sorbian, a possessive adjective can control the agreement of an attributive modifier, as in this example (Corbett 1987:303):

(2) moj-eho muž-ow-a
my-SG.MASC.GEN husband-POSS-SG.FEM.NOM
sotr-a
sister-SG.FEM.NOM
'my husband's sister'

The possessive adjective 'husband's' agrees with the head noun 'sister' showing nominative singular feminine agreement. The interesting part is the attributive modifier 'my' which is masculine, agreeing with the root of the adjective 'husband's' rather than with the head of the noun phrase.

Agreement is currently a live topic. Unfortunately, the terminology is not consistent across frameworks. Some use the term 'agreement' to cover feature matching in a range of domains, from within the noun phrase to antecedent-anaphor relations. Others limit it more or less drastically. Rather than drawing a strict boundary on what is agreement and what is not, we adopt a broader perspective in the database and distinguish between more and less canonical cases of agreement (Corbett, forthcoming). This way, users of different perspectives can be aware how the data relate to their own conception and analyses of the area.

The paper is organised as follows. Section 2 describes the database. Section 3 shows how on the basis of the kind of information that is included in the database a typology of agreement can be constructed. This is followed by a discussion of the implications of our findings both for linguistics and computational linguistics in Section 4.

2. The Agreement Database

The agreement database is a novel sort of typological database in that it includes detailed information on a small, carefully chosen set of languages. Fifteen languages, taken from different families so as to maximise diversity, are being investigated. These are Basque, Chichewa, Georgian, Hungarian, Kayardild, Mayali, Ojibwa, Palauan, Qafar, Russian, Tamil, Tsakhur, Turkana, Yimas, and Yup'ik.[1] For each of these languages, data is gathered according to a consistent format, which is described in Section 2.1, and entered into a relational database for searching and further reference. The structure of the database is described in Section 2.2.

2.1. Data Format

In our database, we use the following framework to describe agreement. We call the element which determines the agreement (e.g. the noun phrase as in the English example above) the controller. The element whose form is determined by agreement is the target (e.g. the verb). The syntactic environment in which agreement occurs is the domain of agreement (e.g. subject-predicate). And when we indicate in what respect there is agreement, we are referring to agreement categories (e.g. number, person). Finally,

[1] This is therefore a different enterprise as compared with the extensive database compiled by Anna Siewierska. By the time of the writing of Siewierska (1999), her database included 272 languages, which is an effective size for checking cross-linguistic claims. Naturally, the information on individual languages is less detailed than in the database that we are constructing.

there may be conditions on agreement (e.g. definiteness). Our framework of terms is illustrated in Figure 1.

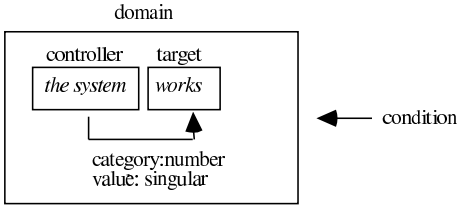

Figure 1: Framework of terms

Each of those areas has been further investigated in our database. Thus for each language we have defined its agreement domains with the respective controllers, targets, categories and their values, and conditions if present. For example, for Chichewa the following agreement domains are distinguished:

Agreement Domain	Frequency
Subject-Predicate	12
Antecedent-Anaphor	6
Head-Modifier	6
Possessor-Possessed	1

The most frequent agreement domain in Chichewa is Subject-Predicate. It occurs with twelve different controller-target pairs. For each of those the database contains a hyperlink to a file with example sentences illustrating the particular kind of agreement. For instance, the following Chichewa sentence illustrates subject-predicate agreement of the noun phrase subject with the finite verb in number:[2]

(3) tsamba li-ku-bvunda
G3SG.leaf G3SG-PRES-rot
'The leaf is rotting'

For each language in the database, there is also a prose report written by the researcher who established and entered the data for that language, giving sources. Data for the different languages is obtained from published grammars of sufficient detail and through personal communication with experts. The reports are intended to allow the user to see which analytical decisions were made and to treat the data accordingly. Since considerable disagreement exists in the literature, it is important that the user can see how choices were made.

2.2. Structure of the database

The agreement database was designed and implemented in Microsoft Access'97. The database contains 9 tables: `Language`, `LanguageDomain`, `Domain`, `ControllerCategory`, `TargetCategory`, `Construction`, `Controller`, `Category`, and `Target`. The relationships between these tables are illustrated in Figure 2.

[2]G3 stands for gender 3. The traditional Bantu concord subclasses are organised into 10 genders in the database.

The tables towards the right of this figure contain the basic elements that we distinguish to define agreement, i.e. possible controllers, possible targets, possible domains, and possible categories. Both controllers and targets can exhibit agreement categories. The category values are not necessarily the same. For example, in British English you can say 'The committee have decided' where a singular controller ('the committee') has plural agreement. The information about the combination of controllers and agreement categories and targets and agreement categories is contained in the tables `ControllerCategory` and `TargetCategory`. The next table to the left is the `Domain` table. It defines unique combinations of agreement constructions with controller (category) and target (category) pairs. Each of these agreement domains is assigned a unique arbitrary index. The `Domain` table is linked to the `LanguageDomain` table which forms the heart of the database. This table combines information about languages with that about agreement domains. The relationship between the `Domain` and `LanguageDomain` table is one-to-many as a particular agreement domain can occur in more than one language. The `LanguageDomain` table is also linked to the `Language` table by a one-to-many relationship as a particular language can have several agreement domains. The `Language` table contains information about the languages in the database. It defines the language family to which a language belongs and there is a hyperlink to the language report.

3. Towards a typology of agreement

The agreement database contains a wealth of information about agreement in a small set of genetically unrelated natural languages. In this way it provides insight into agreement controllers, targets, domains, categories, and conditions for each of the languages individually as well as across languages. It allows us to extract information about them separately (such as their frequency) as well as about how they can be combined. The database tells us what the possible controller-target pairs are, in which domains they occur, which categories occur with these controller-target pairs, etc. So far, the database contains detailed descriptions of the following languages: Basque, Chichewa, Georgian, Hungarian, Mayali, Tamil. The data of five more languages has been prepared. For the six languages in the database, we have identified the following agreement domains: Subject-Predicate, Direct Object-Predicate, Indirect Object-Predicate, Head-Modifier, Possessor-Possessed, Antecedent-Anaphor, Imperative, Allocutive, Pro-forms, and Weather Predicate. Subject-Predicate is the most common agreement domain. All six languages in the database have Subject-Predicate and Antecedent-Anaphor agreement. Direct Object-Predicate and Head-Modifier agreement occur in three of the six languages, whereas Imperative, Indirect Object-Predicate, and Possessor-Possessed agreement occur in two of the six languages. The domains Allocutive, Pro-forms, and Weather Predicate only occur once.

Thirty-five different controller-target pairs have been determined. The most common controller-target pairs are Personal Pronoun-Finite Verb and Noun Phrase-Finite

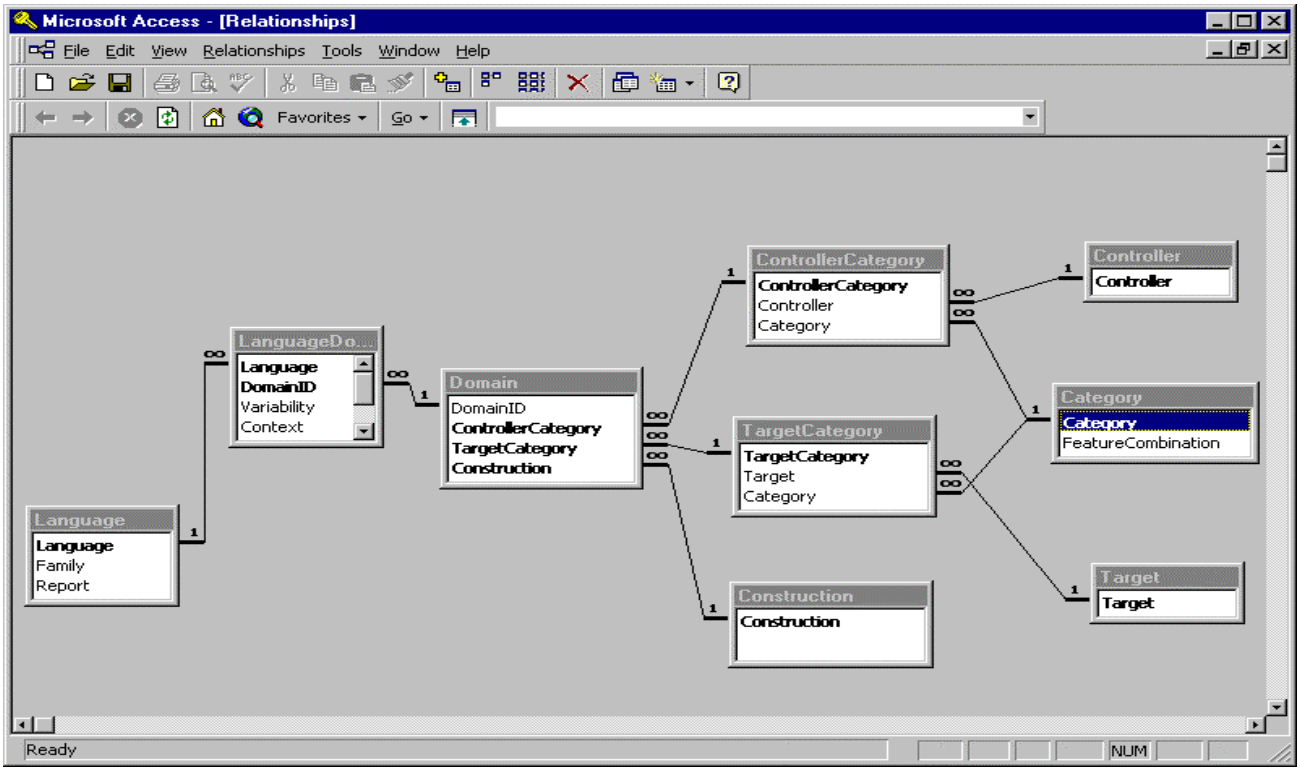

Figure 2: Database Relationships

Verb. Figure 3 gives us an indication of the potential combination of domains with controller-target pairs. We see that the Subject-Predicate domain potentially has the most differentiation of controller-target pairs, allowing for twenty-two different pairs. We would expect less common domains to have less differentiation.

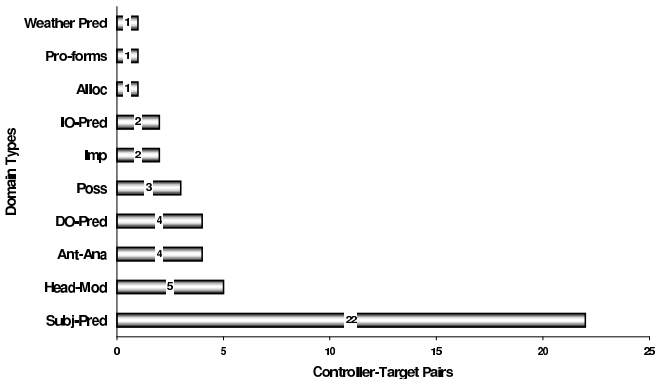

Figure 3: Domains with Controller-Target Pairs

The controller-target pairs were counted independently of the languages and the categories that are involved. The most common agreement categories in the database are number, gender, and person.

If we combine the above information about agreement domains and controller-target pairs with language, we see that Basque shows agreement in the widest variety of domains, i.e. six out of the ten domains that are distinguished so far, but it exhibits relatively little agreement within these domains, the total number of agreement domains with unique controller-target pairs being sixteen in Basque. This amounts to an average of 2.67 different controller-target pairs per domain. Tamil, on the other hand, has only three agreement domains, but relatively large variation within these domains. The total number of agreement domains with unique controller-target pairs is the same as in Basque, i.e. sixteen. This means that Tamil has an average of 5.3 different controller-target pairs per domain. Our database is particularly well-suited to extract this kind of fine-grained information about agreement in the different languages.

Examples of conditions that have been identified in the database are animacy, word order, definiteness, and tense. Figure 4 illustrates the potential domains for each type of condition which we have found in our database. The conditions are counted independently of language to determine how many agreement domains a condition may operate in. For instance, the figure shows that animacy has the potential to be a condition for agreement in at least three different types of agreement domain, whereas definiteness has been found in only one so far. It is important to note that in

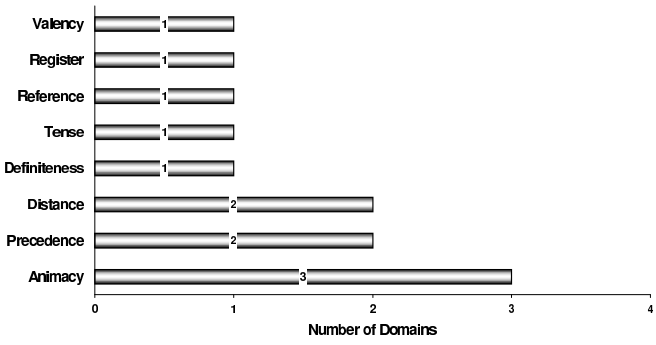

Figure 4: Conditions and Domains

our database conditions on agreement contrast with agreement categories. However, a potential user who, despite the evidence, wishes to treat conditions on agreement with agreement categories, may do so.

Work on the database is ongoing and predictions will be of greater value once all the data has been entered. However, as we saw in this section certain trends can be distinguished.

4. Implications

The database is both of theoretical and practical importance. It is of theoretical importance as the information in the database increases our understanding of natural language syntax and category systems. As we saw in Section 3, it is a step towards the development of a general typology of agreement systems which predicts what is, and what is not, a possible agreement system in a human language. This is the main goal of the database. The database also has practical advantages. Information about the commonalities of the different variables involved in agreement can be used to inform the construction of multilingual NLP systems dealing with agreement. In order to increase reusability and extendability, such systems must deal with what is most common first, what is less common can be added later for specific languages.

5. Summary

In this paper we have described the construction of a typological database of agreement. Our intention is to create a qualitative database where the structure of the database reflect the typology, but allows users of different theoretical perspectives to recover any information which is important for them. This is an ambitious undertaking and it obviously requires detailed description. Therefore we are concentrating on a small number of diverse languages. So far the database contains detailed descriptions of six challenging languages: Basque, Chichewa, Georgian, Hungarian, Mayali, and Tamil. Ultimately the database will include data of fifteen languages.

6. Acknowledgements

The research reported here was supported by the ESRC under grant number R000238228. This support is gratefully acknowledged.

7. References

Corbett, Greville G. (1987). The morphology/syntax interface: evidence from possessive adjectives in Slavonic. *Language* 63, 299-345.

Corbett, Greville G. (forthcoming). Agreement: Canonical instances and the extent of the phenomenon. In Janet DeCesaris, Angela Ralli & Sergio Scalise (eds.) *Proceedings of the Third Mediterranean Morphology Meeting*, Barcelona 2001.

Siewierska, Anna (1999). From anaphoric pronoun to grammatical agreement marker: Why objects don't make it. In G. Corbett (ed.) *Folia Linguistica XXXIII/2. Special Issue on Agreement*, 225–251.

Steele, Susan. (1978). Word order variation: a typological study. In Joseph H. Greenberg, Charles A. Ferguson and Edith A. Moravcsik (eds.) *Universals of Human Language: IV: Syntax*, pages 585–623. Stanford: Stanford University Press.