Constructing a typological database for inflectional morphology:  
the SMG database for syncretism

Dr Dunstan Brown  
Surrey Morphology Group,  
Linguistic and International Studies,  
University of Surrey  
Guildford, Surrey, UK, GU2 7XH  
d.brown@surrey.ac.uk  
www.surrey.ac.uk/LIS/SMG/

Abstract

Typological databases tend to register broad generalizations about languages (e.g. word order). Inflectional morphology may not appear to be readily accessible to database oriented study, because of its apparent language specificity. We discuss three important issues for constructing a database for a morphological phenomenon, namely syncretism: terminology, coverage and structure.

The first issue is terminology. A user needs to know what it is that they are getting. Syncretism, for example, is defined along the lines of Matthews (1997: 367): "The relation between words which have different morphosyntactic features but are identical in form." Furthermore, if one has an empirical orientation, it should be required that a language is defined as syncretizing features, when those features occur elsewhere in that language with morphological consequences. Other uses of the term 'syncretism' treat a language as syncretizing features, precisely because they are not present in that language. The latter definition would make it difficult to create a database, as one would need a list of potential feature values a priori. Furthermore, these are two definitions of quite different things, so it is important that any typological database should come with a detailed definition. This explicitness increases the database’s value for other researchers, and is important for others who may aggregate information from a number of databases.

For coverage there is a trade-off between quality and quantity. Broad coverage may mean that quality of analysis suffers, with no guarantee that similar phenomena are being investigated cross-linguistically. Hence for inflectional phenomena, such as syncretism, it is better to start with a smaller sample of diverse languages, but to be sure that there is a great deal of detail. We have constructed a detailed relational database of inflectional syncretism from 30 languages. The aim has been to enter every type of syncretism found within each of the chosen languages.

The structure of the database may reflect a particular theory for practical reasons. There are a variety of theoretical approaches to syncretism: underspecification, referrals (syncretism as a binary pair of morphosyntactic combinations) and indexing (syncretism as a set of morphosyntactic combinations). Underspecification requires a specific interpretation and may fail to encode all of the information required for searching the database, but there may be little to choose between the referrals and indexing approaches. The structure of our database reflects the referrals approach. Hence, we have developed a typological database where the phenomenon is clearly defined, and the data structured such that a potential user knows what they are getting.
Introduction

This paper discusses the Surrey Morphology Group’s experience of creating a database for inflectional syncretism. In particular, three issues are addressed: in §1 the definition of the phenomenon in question; in §2 the choice of the language sample; in §3 the extent to which the structure of the database reflects particular linguistic theoretical assumptions about morphological syncretism.

Our aim in constructing the database was essentially twofold. We wished to create a resource which would be useful for our detailed investigation of inflectional syncretism, and, in addition to this, make this resource available at some point to the wider community. For this, we would need to be relatively explicit about our use of the term ‘syncretism’ and also, because it is not really possible for data to be entered into a database in an entirely ‘theory-neutral’ way, to provide a potential user of the database with the chance to check how we had come to our analyses. As a result, the theory which underlies the database may be more inclusive than the one which the researchers prefer.

1 Terminology

A typical definition of inflectional syncretism is that of Matthews (1997: 367):

"The relation between words which have different morphosyntactic features but are identical in form."

(Mathews 1997: 367)

If we consider the Classical Armenian u-stem noun ʒam 'time' in table 1, we see that the morphosyntactic feature combinations locative singular, dative singular, genitive singular and instrumental singular in table 1 correspond to the single inflected form ʒamow.

<table>
<thead>
<tr>
<th></th>
<th>u-stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>sg nom</td>
<td>ʒam</td>
</tr>
<tr>
<td>sg acc</td>
<td>ʒam</td>
</tr>
<tr>
<td>sg loc</td>
<td>ʒamow</td>
</tr>
<tr>
<td>sg dat</td>
<td>ʒamow</td>
</tr>
<tr>
<td>sg gen</td>
<td>ʒamow</td>
</tr>
<tr>
<td>sg instr</td>
<td>ʒamow</td>
</tr>
<tr>
<td>sg abl</td>
<td>ʒamè</td>
</tr>
</tbody>
</table>

Table 1: An instance of syncretism in Classical Armenian

Imagine that a potential user of a database on inflectional syncretism had typed in a query to the database and asked to be shown all instances where there was identity between dative singular, locative singular and genitive singular. Among other things, the query then gives the syncretism from table 1. A user would need to know on what basis it was decided that there is syncretism of these forms in Classical Armenian.

Some uses of the term ‘syncretism’ treat it in a very broad way, such that it is not established on a language internal basis, but relies on global knowledge of the inventory of features. So dative, locative and instrumental would be instances of a syncretism, because the language in question always fails to have them. There are a number of problems with this: i) we would need to know the inventory of features in advance; ii) determining which functions the totally absent function is syncretic with usually involves the assumption of prototypical semantics for related functions. Here, however, the definition of the phenomenon may already involve a particular assumption about why it comes about. In this case, a definition of the phenomenon in question is ultimately based on the view that its causes are underlyingly semantic (see §3.1). But within the theoretical space of possibilities we might consider a number of motivations for syncretism: semantic, syntactic, morphological, phonological. A database should not be biased toward any of
these possible interpretations, and should ideally allow a researcher to discover possible motivations. Construction of a database on syncretism where the phenomenon was not defined in a language-internal way would, for the reasons given, obscure or hide morphological and phonological causes, for example.

Accordingly, we need to refine the definition such that syncretism is an instance of a single inflected form corresponding to more than one morphosyntactic description, where feature values in the morphosyntactic description are established on a distributional basis within the language. Comrie (1991:46) makes a distinction between distributional and formal case, for instance.

“For each nominal in the language, establish the distinct forms that this nominal can show [i.e. its array of formal cases]. Now compare the distributions of all nominals. If some distribution is of a distinct form for all nominals, then this is a [distributional] case. If the distribution (a) of some form of some nominal is a proper subset of the distribution (a+b) of some form of any nominal, then the distribution or subdistribution defined by a and b are distinct [distributional] cases for all nominals. If the distribution (c+d) of some form of some nominal mutually and nonexhaustively overlaps the distribution (d+e) of some form of any other nominal, then each of c, d, and e is a distinct distributional case for all nominals.”

(Comrie 1991:46)

Hence, for a particular feature value, in this instance a value for case, to be involved in syncretism and therefore be entered into the database, the feature-value must formally be distinguished somewhere within the language in question. On the basis of this we can define syncretism informally, but relatively explicitly, as follows:

**Syncretism**
A single inflected word form corresponds to more than one morphosyntactic description, where the feature values in the descriptions are established distributional values for the language in question.

So locative and instrumental are treated as valid cases for Classical Armenian, because they have distinct forms for some nouns (table 2).

<table>
<thead>
<tr>
<th>Feature</th>
<th>a-stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>sg nom</td>
<td>gir</td>
</tr>
<tr>
<td>sg acc</td>
<td></td>
</tr>
<tr>
<td>sg loc</td>
<td>giri</td>
</tr>
<tr>
<td>sg dat</td>
<td>giroy</td>
</tr>
<tr>
<td>sg gen</td>
<td></td>
</tr>
<tr>
<td>sg instr</td>
<td>giroy</td>
</tr>
<tr>
<td>sg abl</td>
<td>giroy</td>
</tr>
</tbody>
</table>

Table 2: Classical Armenian ‘book’

Equally, in the demonstrative and pronominal system there are distinct genitive singular forms. So the use of a formal and distributional definition yields clear instances of the phenomenon on a language-internal basis.

We have illustrated here with examples of case (in the singular), but under the definition of syncretism given, we can include any morphosyntactic features.1 Within our database we have defined ten possible feature sets, together with provision for two additional sets, when required. The sets are Number, Case, Gender, Definiteness, Person, Tense, Mood, Voice, Aspect. Negation. As we have pointed out, we did not have an a priori list of the possible feature values for these sets. Instead, we designed the database in such a way that the researcher could enter new values when appropriate.

2 Coverage

Given the definition of syncretism in §1, populating the database involves examination of

---

1 See, for example, the complex person syncretisms in Dalabon (Evans, Brown & Corbett 2001).
the paradigms of related sets of lexical items and comparing them with the paradigms of other lexical items of the same or related word classes. This obviously requires grammatical sources which are quite detailed in their description of the language’s morphology. Typically, studies of syncretism have been biased toward Indo-European, despite the fact that it is found elsewhere.

The Surrey Morphology Group has carried out larger scale surveys of Person and Case syncretism as part of the World Atlas of Linguistic Structures project (Baerman and Brown in progress a, in progress b). However, trying to do a detailed analysis of every type of syncretism within a language requires high quality data which may not be available for a larger sample.

<table>
<thead>
<tr>
<th>Family</th>
<th>Number of Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-European</td>
<td>4</td>
</tr>
<tr>
<td>Afroasiatic</td>
<td>2</td>
</tr>
<tr>
<td>Niger-Kordofanian</td>
<td>2</td>
</tr>
<tr>
<td>Nilo-Saharan</td>
<td>2</td>
</tr>
<tr>
<td>Tibeto-Burman</td>
<td>2</td>
</tr>
<tr>
<td>Trans-New Guinea</td>
<td>2</td>
</tr>
<tr>
<td>Altaic</td>
<td>1</td>
</tr>
<tr>
<td>Austronesian</td>
<td>1</td>
</tr>
<tr>
<td>Carib</td>
<td>1</td>
</tr>
<tr>
<td>Chibchan</td>
<td>1</td>
</tr>
<tr>
<td>Chukotko-Kamchatkan</td>
<td>1</td>
</tr>
<tr>
<td>Dravidian</td>
<td>1</td>
</tr>
<tr>
<td>Eskimo-Aleut</td>
<td>1</td>
</tr>
<tr>
<td>Isolate</td>
<td>1</td>
</tr>
<tr>
<td>Kartvelian</td>
<td>1</td>
</tr>
<tr>
<td>Kiowa-Tanoan</td>
<td>1</td>
</tr>
<tr>
<td>Nakh-Daghestanian</td>
<td>1</td>
</tr>
<tr>
<td>Non-Pama-Nyungan</td>
<td>1</td>
</tr>
<tr>
<td>Pama-Nyungan</td>
<td>1</td>
</tr>
<tr>
<td>Sepik-Ramu</td>
<td>1</td>
</tr>
<tr>
<td>Taconan</td>
<td>1</td>
</tr>
<tr>
<td>Uralic</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 3: Genetic Affiliations in the Database

For our purposes data collection was guided by the desideratum of instantiating as much as possible of the logical feature space. Data was obtained from published grammars of sufficient detail and through personal communication with experts. This meant that we limited ourselves to 30 languages. In table 3 we have shown how many languages there are from each genetic group.

The database contains 1256 records on languages, syncretisms and their domains (counted from the LanguageSyncretismDomain table, see §3.2). That is on average 42 records per language, although some languages have much more information associated with them than others.

In sum, rather than concentrate on a broad sample with little information for each language, we have created a smaller scale detailed database to examine the logical space of syncretism using high quality grammars.

3 Structure

The structure of the database may reflect a particular theory for practical reasons.

3.1 Linguistic Theory and Data

From a theoretical viewpoint there are at least four ways in which inflectional syncretism can be interpreted. The first is to claim that it is accidental homophony and should be interpreted as a mere coincidence of form (i.e. the forms just happen to be the same, but are not identical). The second is to treat the syncretism as an instance of where one morphosyntactic combination borrows its form from another (typically called referrals). The third way is to claim that particular morphosyntactic combinations share a common form to which they are both indexed (i.e. one does not borrow from the other). The fourth way involves some reinterpretation of morphosyntactic features: typically, it is assumed that, where two or more functions share the same form, there is some degree of underspecification involved. The assumption is that morphosyntactic features are themselves defeasible.

2Matthew Baerman has created a database specifically for Person syncretism with a coverage of 1101 languages.
In its extreme form the first approach basically says that the phenomenon does not exist. Of course, there are variants on the first approach which treat certain apparent identities as accidental, and others as not. The problem with the first interpretation of syncretism is that it is often very difficult to determine what is accidental and what is not. For instance, in Russian adjectives the (masculine and neuter) instrumental singular and dative plural always have the same form, a fact which would suggest that it is hardly accidental, and yet this may appear counterintuitive to proponents of particular theories. Because it is a matter of debate what is accidental and what is not, it is important that every instance of syncretism is entered in the database. It should be the decision of the database user whether a syncretism is accidental or not. In order to allow for this, however, we must build in a means for the user to check the data. As we shall see in §3.2 this is achieved by the use of two hyperlink fields in our database (one in the table of languages and one in the table which combines languages and syncretisms).

Irrespective of its theoretical merits, the fourth approach, using underspecification, is unsatisfactory for the construction of a database, as it involves a specific interpretation of the data. In addition to this, there are two problems. We might start to construct a database and find over time that the feature ‘geometry’ we have assumed does not function for each new language we come across. Whether it would or not is an empirical matter, of course, but it cannot be guaranteed from the outset. Another potential problem is that it will leave out information that a potential user would find useful. In other words, the user would have to reconstruct the analysis in order to determine which features are syncretic. We have argued elsewhere that some of the assumptions associated with the relationship between syncretism and underspecification approaches do not necessarily apply over a broad sample of languages (Baerman, Brown, Corbett in press). Irrespective of the truth of this claim, use of an underspecification based approach to create a database of syncretism omits information that will be useful for a potential user.

Having excluded on practical grounds the first and fourth theoretical approaches as a basis for the structure of our database, we now turn to the referrals and indexing approaches. These approaches are better, because they need not force a particular interpretation of the data. The difference between the third and fourth approach is more one of implementation than theory. Consider the Russian paradigm in table 4.

<table>
<thead>
<tr>
<th></th>
<th>zakon</th>
<th>komnata</th>
<th>kost’</th>
<th>okno</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg Gen</td>
<td>zakona</td>
<td>komnati</td>
<td>kost’i</td>
<td>okna</td>
</tr>
<tr>
<td>Sg Dat</td>
<td>zakonu</td>
<td>komnate</td>
<td>kost’i</td>
<td>oknu</td>
</tr>
<tr>
<td>Sg Loc</td>
<td>zakone</td>
<td>komnate</td>
<td>kost’i</td>
<td>okne</td>
</tr>
</tbody>
</table>

**Table 4:** Russian syncretisms (in transcription)

One way of accounting for the syncretisms in table 4 is to make appeal to rules of referral. These are rules which specify that one morphological form will be realized identically to another; the term is due to Zwicky (1985: 372). Rules of referral may be seen as comparable to Perlmuter and Orešnik’s (1973) ‘prediction rules’. Stump (1993, 2001) argues explicitly that underspecification can account for certain syncretisms, whereas referrals are the best way of dealing with others. Under the referrals approach we would say that the form of the dative singular is based on the locative singular (because the locative singular form ending in –e can be found on its own in the zakon and okno declensions). If we stated that there is a referral of dative singular and locative singular in Russian, this abstracts away from the particular forms, as in the kost’ type the actual form involved differs from that in the komnata type.4 So for Russian we have a syncretism stated as one binary pair of combinations (namely dative singular and locative singular).

---
3 Russian is not one of the languages in the SMG database. We have tried to minimize the number of Indo-European languages.
4 In DATR treatments of morphology using default inheritance this kind of abstract relationship can be represented by ‘global inheritance’, for example (Evans and Gazdar 1996). In default unification approaches it is represented using reentrancies which may be indefeasible or default (for example, Lascarides and Copestake 1999).
For the *kost* type we also need to say that the genitive singular is syncretic with both the dative and locative singular. Of course, by stating that it is syncretic with either one should involve the implication that it is syncretic with the other. However, for practical reasons this information may not be apparent to the user of a database. As we shall see in our discussion in §3.2 we introduce a particular distinction between ‘dependent’ and ‘independent’ syncretisms to get round this issue.

Treating the forms in table 4 as indexed to particular morphosyntactic combinations is another way that could be used to represent the syncretisms in table 4. Practically there is a difference from the referrals approach. Here we would say that there are two types of syncretism in table 4: one (*komnata* type) involves the set of feature combinations *SgDat*, *SgLoc*: the other involves the set of feature combinations *SgDat*, *SgLoc*, *SgGen*. This approach is feasible, although implementing it would probably require some estimate on the upper limit of the number of combinations in a set. The practical issue for constructing a database using this approach is the use of values which enable us to search and find instances where *SgDat* and *SgGen* are syncretic, for example. Under the referrals approach we can do this straightforwardly, because we are always dealing with instances of binary pairs of feature combinations. The indexing approach can sometimes lose sight of what is shared. For example, the syncretism of *SgDat* and *SgLoc* is shared by both the *komnata* and *kost* types, but they are treated as separate sets - one subsuming the other - under the indexing approach. It is also desirable to constrain the data entry such that we can use such constraints as primary keys to say that once we have stated that a language has a particular syncretism (irrespective of the form) we do not repeat this. It is for these reasons that we adopted the ‘referrals’ approach to the design of our database.

Aronoff (1994: 83) criticizes the use of rules of referral - in certain analyses - precisely because of their directionality. In our implementation, no directionality is implied, other than to say that there are identical forms used by the two feature combinations in question.5

Using an approach based on binary pairs of combinations still allows us to search for hierarchical ordering in the data. In the work of Hjelmslev (1943/1961), Carstairs-McCarthy (1984: 1987 and 1992), Wunderlich and Fabri (1995), and Noyer (1997: xx-xxi, 45), for example, particular categories (feature sets) are ordered in relation to each other to capture dependencies, such as that between gender and number. For instance, Gender distinctions, for example, are lost in the presence of a particular number, typically plural (Greenberg 1963). Treating syncretisms as binary pairs of feature combinations allows us to look at the context in which they occur. Taking up the example of case syncretism within a particular number, we can easily construct queries in which we ask for instances where the number values are the same, but the case values differ. Hence, the adopted approach still allows for this interpretation, and it is possible to search for data which may indicate such hierarchical orderings.

### 3.2 The Database

The Surrey Morphology Group Syncretism Database contains 18 tables. Of these, 10 tables contain the values for reasonably well-defined feature sets. Hence, there is a table Number which contains the values for number the researcher finds as he enters data from a chosen language. The 10 tables are: Number, Case, Gender, Definiteness, Person, Tense, Mood, Voice, Aspect, Negation. In addition to the 10 tables there are two others, Spare1 and Spare2 for additional morphosyntactic feature values which cannot be placed in any of the sets given in the 10 tables. These 12 tables are to be seen on the right side of figure 1.

---

5 As two feature combinations are involved we have to use an arbitrary convention to decide which one is on the left-hand side of the binary pair. But this has no theoretical import.
Figure 1: Database Relationships

A feature value from one of the twelve tables can be combined with other feature values from the same twelve to form a morphosyntactic combination. The Combination table is ringed in figure 1. The relationship between any given feature table (e.g., Number) and the Combination table is one-to-many. The Combination table assigns a unique arbitrary index to the combination set, but it is the actual set of combinations which is the primary key for the Combination table.

The table of syncretisms is in the middle of the top set of tables. The Combination table has two relationships with the Syncretism table. Both of these are one-to-many. A syncretism is treated as a binary pair of morphosyntactic combinations. The Syncretism table contains the indices for each morphosyntactic combination. These indices are automatically compared during data entry to check that they are not identical (as a form which shares identical morphosyntactic features is not a syncretism) and to determine which combination is put in the first field, and which in the second. This is done by placing the higher number in the second field. The index number for the combination is determined purely by the particular point in time at which the researcher entered the combination. If one wishes to compare all instances of, say, pairs of number values involved in syncretisms, irrespective of the other features they occur with, this may mean that queries have to reorder the values alphanumerically. Consider the following pairs of combinations taken from the database:

(1) Combination 1
   pl-masc
   Combination 2
   sg-fem

(2) Combination 1
   sg-masc
   Combination 2
   pl-fem

In both these examples the number values involved are sg and pl. However, in (1) pl is the Combination 1 value, and in (2) sg is the Combination 1 value. Hence, queries which may be interested in pairs of single categories such as Number also involve sorting the values alphanumerically. This has proved to be straightforward. If we had decided to adopt an approach in which forms were associated with an n-tuple of combinations, this ordering issue would be more complex, of course. Other fields in the LanguageSyncretismDomain table (see later) allow us to associate a binary pair with other binary pairs, if they are part of a greater syncretism. We do this by defining a syncretism as ‘independent’ if it may occur on its own (for instance, the SgDat-SgLoc syncretism in Russian is ‘independent’) and

---

6 Figure 1 shows two tables ringed. In fact, these are just the table of combinations repeated. This is because the Combination table has two relationships with the Syncretism table: a combination occurs in both the left-hand side field and right-hand side fields of a syncretism.
‘dependent’ if the binary pair only occurs in the presence of another syncretism (e.g. the SgGen-SgLoc syncretism only occurs in the presence of the SgDat-SgLoc syncretism). A ‘realization’ field then allows us to find examples of syncretisms based on the same form.

Toward the bottom left of figure 1 is a table which contains fields of word classes (nouns, adjectives, verbs etc.). This table can construct sets of word classes to be used in the Domain table. The primary key for this table is the field which contains the value which represents a set of word classes. A set consists of one or more word classes. This then enables us to see which word classes typically group together in syncretism domains. The Word Class table has a relationship with the Domain table, which combines three fields: word class, syntax, semantics. Hence, the relationship between the Word Class table and the Domain table is one-to-many, as a single set of word classes could occur under different syntactic or semantic restrictions.

The table in the top left contains information about languages. There are three fields: Language, Family, Report. Introduction of the Family field enables us to make restrictions on queries where all examples may be from the same language family. Also, given the non-redundant design of the database, we can freely change a family affiliation, if required on the basis of new information, by just changing one field of one record in one table. The Report field is a hyperlink field which contains a detailed report on the language written by the researcher.

To its right the Language table has the Language Syncretism Domain table (LSD table), which is the heart of the database. It combines the information about languages with that about syncretisms and domains. The relationship between the Language table and the LSD table is one-to-many. The relationship between the Syncretism table and the LSD table is one-to-many, and the relationship between the Domain table and the LSD table is also one-to-many. The LSD table also contains a hyperlink field which links to a document containing illustrative paradigms for each syncretism in the database. This thereby allows for quality control by the user, who can see how the analysis, encoded by the choice of morphosyntactic combinations, is arrived at.

Conclusion

In §1 we gave an explicit definition of syncretism and showed how important this is in order to guarantee the grounds of comparison. We then discussed theoretical issues and showed how particular theories are better suited for database design for practical reasons. Finally, in §3 we showed the design of our database and how the structure reflected, to a certain extent, one particular theoretical approach, at the same time allowing a user access to all the information they require to check analyses.

Acknowledgements

The research reported here was supported by the Economic and Social Research Council (UK) under grant number R000237939. Their support is gratefully acknowledged. The work presented here benefits from discussions within the Surrey Morphology Group. I would especially like to thank Matthew Baerman, who is responsible for data analysis and entry, and whose ideas, suggestions and work have informed the development of the database at each stage. I am grateful to Greville Corbett for discussion of numerous aspects of syncretism and the theoretical issues surrounding it. I would also like to thank Roger Gentry for his help during the design and implementation phase, and David Bowers for initial discussions during the design.

---

7 The set of word classes is labelled WordType. Individual word classes are labelled WordType1 and so on.

8 We have found that there are very few instances of this. A potential example is the word class (set) adjective, which may behave differently in predicative or attributive syntactic function. But empirically it appears that examples like this are rare. Hence, use of the database has shown that a theoretical possibility is not that common.
References


