

Formal Approaches to Slavic Linguistics¹

Greville G. Corbett
University of Surrey, UK

1. Introduction

This paper is intended to celebrate the FASL series of conferences, by reflecting on the fine choice of title. I imagine we do *Linguistics* because we find language fascinating and believe that, as technical means of communication become ever more available, issues of the use of those channels, and specifically communication through language, will in turn grow in importance. For the value of the *Slavic* contribution to linguistics one need only think of Jakobson and Trubetzkoy. However, those who set up the series might have been tempted to focus it on Russian. While Russian has a dominant position, given its status as a world language and hence its role in educational institutions, it was so much better to have *Slavic Linguistics* as the subject. The linguistic interest of Slovene and Sorbian for instance is just as great as that of Russian. So to the more substantial issue, that of *Formal Approaches*. There is a variety of formal approaches which may be of benefit for Slavic linguistics. Some have been discussed at previous conferences, and the current collection is refreshingly diverse in this respect. I will outline three different approaches. In each I will report on joint work with several colleagues and highlight the importance of Slavic data for wider typological concerns. The three approaches involve morphology, lexical semantics and corpus linguistics. Since I wish to illustrate breadth I shall not be able to cover each approach in depth, rather I shall give illustrations with pointers to fuller accounts.

¹The support of the ESRC UK (grant R000237845) is gratefully acknowledged. I also wish to thank my co-researchers. This draft was improved following suggestions by Andrew Hippisley, lively discussion at FASL 8, and helpful comments by the editors.

2. Network Morphology

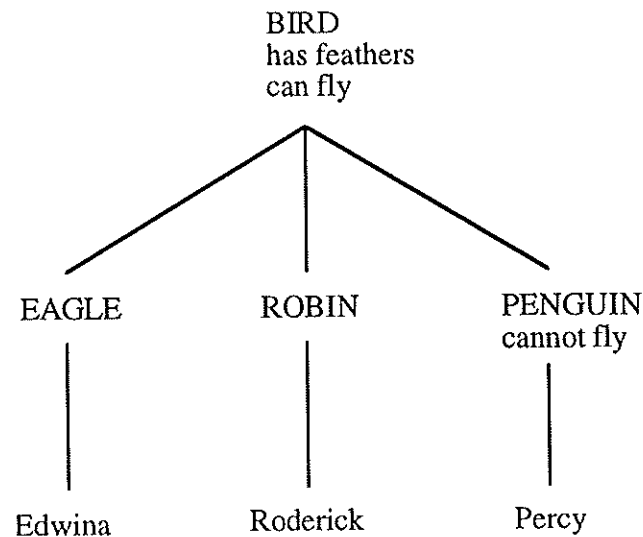
The first formal approach is Network Morphology. I will introduce some basic ideas and then consider two illustrative examples, gender assignment and syncretism. The research to be discussed also involves Norman Fraser and Dunstan Brown, and the account draws on previous publications. There is space here for a sketch; details are given in Corbett & Fraser (1993, 1997), Brown & Hippisley (1994), Fraser & Corbett (1995, 1997), Brown, Corbett, Fraser, Hippisley & Timberlake (1996), Hippisley (1997), Brown (1998).

Network Morphology is an approach to morphology which distributes information across a network in which generalizations can be optimally expressed. Generalizations become available in specific cases by the operation of default inheritance. Network Morphology theories are expressed in a formal representation language called DATR developed by Roger Evans and Gerald Gazdar. DATR is a particularly useful formalism for developing explicit accounts of complex linguistic data because it is formally well-defined (Evans & Gazdar 1989a) and it allows for the construction of largely declarative accounts which rely on a limited set of basic operations, of which default inheritance is one (Evans & Gazdar 1989b); it has been a major source of inspiration in the development of Network Morphology. Helpful introductions to DATR for linguists can be found in Evans & Gazdar (1996) and Gazdar (forthcoming). Added to the formal rigour and rich expressiveness of the DATR language is a third valuable feature: computer interpreters (and compilers) exist which are capable of taking a linguistic theory expressed in DATR as input and automatically generating as output all the forms which the theory allows. Working in this way means taking seriously some of the basic ideas of generative grammar. The computer thus has a valuable checking role: we are interested in *theoretical* linguistics, being concerned with observation, description and explanation rather than computational issues like algorithmic efficiency. We wish to separate the questions of linguistic theory (Network Morphology) from the formalism we use (DATR). This separation focuses

attention on the substance of theories rather than on their notation (Shieber 1987).

We should first consider the concept of default inheritance, which we approach using the taxonomic hierarchy in Figure 1. The lines in the taxonomy indicate instantiation rather than sub-classification. So an eagle is a bird, as is a robin and a penguin; Edwina is an eagle, Roderick is a robin, and Percy is a penguin.

Figure 1. A Simple Instantiation Hierarchy



Given an instantiation hierarchy of this kind, default inheritance allows all attributes of a given node in the hierarchy (such as BIRD) to be inherited by a node which instantiates it (such as EAGLE). This is the case except where the lower node already has a value for some attribute and thus overrides the default (that is, inheritable) value for that attribute. In Figure 1 a BIRD has feathers and can fly. These facts are inherited by EAGLE and ROBIN and, indirectly, by Edwina and Roderick. The attribute of having feathers is also inherited by PENGUIN and, through it, by Percy. However, specific information that PENGUINS cannot fly blocks inheritance

of the more general information about BIRDS. Although Percy is a BIRD, he cannot fly.

Default inheritance allows generalizations to be expressed once at a high level, and then automatically to apply to everything which inherits from there. Regularities, subregularities and exceptions can be represented easily and economically. This approach has the added advantage of marking exceptions as such, as in the case of PENGUIN in Figure 1. If Percy were a penguin who could fly, this extreme exceptionality would be visible because an exceptional fact (PERCY can fly) would override an exceptional fact (PENGUINS cannot fly), overriding a default (BIRDS can fly). Further information on default inheritance can be found in Gazdar (1987) and Daelemans, de Smedt & Gazdar (1992).

The information in Figure 1 could be expressed in DATR as follows:

```
(1) BIRD:
    <has_feathers> == yes
    <can_fly> == yes.
```

```
EAGLE:
    <> == BIRD.
```

```
ROBIN:
    <> == BIRD.
```

```
PENGUIN:
    <> == BIRD
    <can_fly> == no.
```

```
Edwina:
    <> == EAGLE.
```

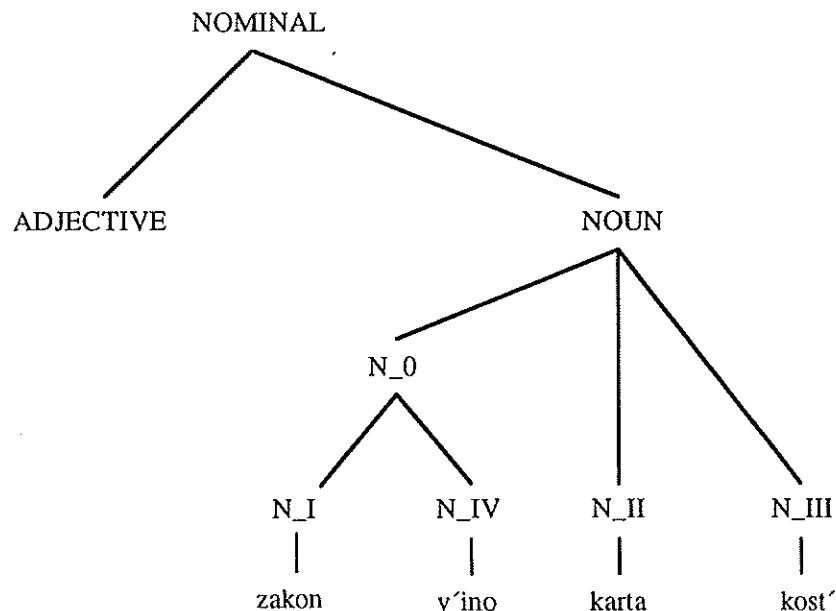
```
Roderick:
    <> == ROBIN.
```

```
Percy:
    <> == PENGUIN.
```

The labels preceding colons are 'nodes'; the angle bracket expressions to the left of the '=' symbol are 'paths'; the words to the right of non-empty paths are 'values'. Thus, the value of the <can_fly> path at the PENGUIN node is 'no'. The form '<>' is the special case in which a path is empty (hence maximally underspecified). This allows, for example, the node 'EAGLE' to inherit all values available at the node BIRD. Since we are dealing with default inheritance, PENGUIN inherits all values from BIRD, except the one which is overridden.

Of course, we are interested in the use of this kind of formalism for expressing linguistic generalizations. Figure 2 shows a simplified inheritance network for the morphology of Russian nominals.²

Figure 2. An Inheritance Structure for Russian Nominals



² Corbett & Fraser (1993): later papers use a network of hierarchies rather than a single one.

N_I to N_IV are nodes where information which distinguishes declensional classes is stored. Notice the posited node N_0 from which N_I and N_IV inherit; it allows us to capture the advantages of analyses of Russian which postulate three noun declensional classes and of those which distinguish four. The following (incomplete) fragment is taken from our earlier analysis (Russian forms in this section are given in phonemic transcription):

```

(2) NOUN:
    <mor loc sg> == "<stem>" _e
    <mor nom pl> == "<stem>" _i.

N_III:
    <> == NOUN
    <mor loc sg> == "<mor dat sg>".

Kost':
    <> == N_III
    <stem> == kost'.
  
```

The first fact at NOUN should be read as saying that the locative singular consists of the stem followed by an *-e* ending. A path enclosed in double quotes in a DATR sentence is used to retrieve the specified value for that path at the node from which the query originates. If we wanted to find the nominative plural of *Kost'*, we would inherit the sentence <mor nom pl> == "<stem>" _i. We would have to find out what the <stem> of *Kost'* is. Since the answer is *kost'*, the nominative plural is *kost'i*. The quoted path means that we take the stem of *Kost'* (not of NOUN, which has no stem). If, however, we wanted to know the locative singular of *Kost'*, we would never inherit the definition of locative singular at NOUN because it is overridden at N_III, from which *Kost'* inherits. The definition of locative singular at N_III establishes an asymmetric identity between the locative singular form of an N_III noun and its dative singular.

It may be desirable to inherit most information from one source, but to have access to some information stored elsewhere. Consider the following fragment:³

- (3) N_II:
 <> == NOUN
 <mor gen sg> == "<stem>" _i.
- N_III:
 <> == NOUN
 <mor gen sg> == N_II.

This says that N_III may inherit its schema for forming the genitive singular from N_II, even though N_III (like N_II) inherits primarily from NOUN. This may be expressed more explicitly as follows:

- (4) N_III:
 <mor gen sg> == N_II:<mor gen sg>.

This was a brief introduction to default inheritance and to the DATR formalism. We now consider two linguistic problems and show how the formal approach sketched combines with a typological approach.

2.1. Gender Assignment

Gender systems have agreement as their defining characteristic. Nouns of a gender language can be grouped analytically according to agreement evidence. We then ask how the native speaker, who produces the agreement evidence, 'knows' the gender of the different nouns. Assignment to a particular gender is always

³ Only facts relevant to the discussion are shown at nodes N_II and N_III in (3). In our full analysis of Russian nominal morphology, each of these nodes records a much richer set of facts. The facts for N_III are quite different from those for N_II, so an analysis of the form 'N_III: <> == N_II' would fail.

possible for the vast majority of nouns, from information required independently in the lexical entry (Corbett 1991:7-69). The particular type of information which may be used gives us a typology of assignment systems. We find *semantic systems* (where only semantic information is required) and *semantic + formal systems* (where semantic information is supplemented by morphological and/or phonological information). Purely formal systems (where gender would be predicted by formal means but where the different agreement classes of nouns would have no semantic significance) are not found.

Languages with semantic assignment and those with semantic rules supplemented by phonological rules are relatively unproblematic. The most difficult are the formal-morphological systems (see, for instance, Aronoff 1994:73-74), and this is precisely what is proposed for Russian, and Slavic more generally. These systems have often been analysed differently; instead of gender being predictable (and therefore not needing to be specified in the lexical entry), some treat gender as specified, and from it attempt to predict the morphological class of nouns. When the number of genders and the number of declensional classes are the same or nearly so, it is not immediately obvious which analysis is to be preferred. I propose that Russian has a gender assignment system in which morphological information supplements semantics (recall that in this section Russian examples are transcribed):

- (5) *Semantic Assignment Rules*
 (5a) Sex-differentiable nouns denoting males (humans and higher animals) are masculine: *sin* 'son', *d'ad'a* 'uncle', *lev* 'lion';
 (5b) Sex-differentiable nouns denoting females are feminine: *doč'* 'daughter', *t'ot'a* 'aunt', *l'v'ica* 'lioness'.

Nouns which are sex-differentiable are those where the sex matters to humans (as for humans and domesticated animals) and where the difference is striking (as in the case of lions). There are few exceptions to these rules but many nouns are not covered by them. Unlike Godoberi, Russian does not treat all noun in the semantic

residue in the same fashion. They are subject to further rules, including the following:

- (6) *Morphological Assignment Rules*
 (6a) nouns of declensional class I (*zakon* 'law' type) are masculine;
 (6b) nouns of declensional classes II (*karta* 'map') and III (*kost* 'bone') are feminine;
 (6c) others are neuter.

Given the dispute as to whether this is the right analysis, there are two traditional types of argument available here. First, and most important, there are language-specific arguments. It can be shown that predicting gender on the basis of declensional class is simpler and involves fewer exceptions than the attempt to predict declensional class on the basis of gender. These arguments are treated at length in Corbett (1982), and will not be repeated here. Second, there is the typological argument: since there are many languages where gender is straightforwardly predictable, it is simpler to claim that it is predictable in all languages, with typological variation being restricted to the type of information used for prediction.

By giving a Network Morphology analysis, using DATR as a tool, we have access to third type of argument: since DATR comes with a compiler, we can demonstrate that our analysis (which is an explicit account of the interactions of semantics, gender and declensional class) does indeed yield the correct results. Given our analysis and the lexical entries for a range of Russian nouns, a computer can be used to verify that our analysis makes the right predictions as to gender. Such an analysis is presented in Fraser & Corbett (1995). The aim of this section is not to justify that analysis. Rather we want to emphasize that the analysis, that of a theoretical linguist working within the Network Morphology framework, can be shown to work using computational methods. Other analyses of gender in Russian are not backed by similar demonstrations of accuracy. Thus formal tools like DATR can elucidate cases which are crucial for typological purposes.

2.2. Syncretism

In our first example, the impetus came from typology for us to provide a formal and hence testable account of critical data. Now we move to an example of the reverse, an attempt at formal description which leads to a typology. The varying patterns of neutralization in Slavic are intuitively of different types, and the differences in the natural ways of handling them in DATR support this view. Some correspond to the notion of 'syncretism', where a single inflected form corresponds to more than one morphosyntactic description (Spencer 1991:45), or informally where the morphology 'fails' the syntax. Work within Network Morphology has led us to a typology of these neutralizations, first according to their domain and second according to their nature.

The Domain of Syncretism. Consider the singular paradigm of *kost* 'bone':

(7)	NOM(inative)	<i>kost</i> '
	ACC(usative)	<i>kost</i> '
	GEN(itive)	<i>kost</i> 'i
	DAT(ive)	<i>kost</i> 'i
	INST(rumental)	<i>kost</i> 'ju
	LOC(ative)	<i>kost</i> 'i

Among other things, we want to say that genitive and dative singular are identical. We could reflect this in the lexical entry for *kost*':

- (8) *Kost*' :
 <> == N_III
 <gloss> == bone
 <mor dat sg> == <mor gen sg>
 ...

However, if we express the identity in this way, it will apply only to the lexical entry for this one word. Such a situation, a syncretism involving a single lexical item, forms the first part of our typology.

However, we are not aware of any instances in Russian. In fact, this syncretism holds for all members of class N_III, the nouns like *kost*. We may position this at a higher point in the inheritance hierarchy, at the node for nouns of declensional type III, from which *kost* inherits:

(9) N_III:
 ...
 <mor dat sg> == "<mor gen sg>"

In both instances the identity is handled under a single node — there is no need for multiple inheritance. The significance of the quotes, as discussed earlier, is that the dative singular is whatever “your own” genitive singular is: in other words dative singular will take the value of genitive singular at the original query node.

The statement of identity as given could be pushed ever higher up the inheritance tree (DATR encourages us to state generalizations higher and higher), and this gives us the first parameter of our typology:

- (10) The Domain of Syncretism
- a single word
 - a single inflectional class
 - a subset of the inflectional classes of a word class
 (In Russian, dative and locative singular are identical in two inflectional classes.)
 - across all inflectional classes in the word class
 (Russian adjectives have genitive plural identical to locative plural.)
 - across more than one word class
 (In Slovene, nominative and accusative dual are identical for all nouns and adjectives, Priestly 1993:399.)
 - across all potentially relevant word classes
 (In Slovene, dative and instrumental dual are identical for anything which can mark them, Priestly 1993:399.)

The Nature of Syncretism. Turning to the nature of syncretism, let us consider the data in Table 1, giving some of the forms of Russian *student* ‘student’ and *zakon* ‘law’ (in transcription):

Table 1. The Morphological Effect of Animacy

	Singular		Plural	
NOM	student	zakon	studenti	zakoni
ACC	studenta	zakon	studentov	zakoni
GEN	studenta	zakona	studentov	zakonov

Let us start with the first column of forms. There is syncretism of accusative and genitive singular (conditioned by whether the noun is animate or inanimate). As we shall see below, this can be captured by a DATR statement including the following:⁴

(11)
 <acc sg animate masc> == "<mor gen sg>"

Basically this is saying that the accusative is the same as the genitive. If we looked at the other paradigms, we would see that they share the same *pattern* of identity, even though the particular inflections differ. It would not be sufficient to state the identity of forms separately for each paradigm; that would imply that the cases involved could equally well differ from paradigm to paradigm. This regularity can be captured in the DATR account by a statement high up the inheritance tree:

(12)
 NOMINAL:
 <acc> == "<mor nom>"
 <acc pl animate> == "<mor gen pl>"
 <acc sg animate masc> == "<mor gen sg>"

⁴ The gender value is available as described in section 2.1.

```
<mor acc $number>
  == < acc $number "<syn animacy>"
      "<syn gender>" >
```

Let us go back to our DATR statement (11), which is embedded in (12). Note that this is not a symmetrical relationship. The genitive form is "right", the accusative is a copy of it. This can be seen by comparing with the second column: any noun of this type will have the genitive singular in *-a*, the accusative matches this genitive if the noun is animate, and the nominative if not. The question of directionality is a live issue. Rules which specify that one morphological form will be realized identically to another are often called 'rules of referral', following Zwicky (1985:372). They may be seen as comparable to Perlmutter & Orešnik's 'prediction rules' (1973). It is precisely because of their directionality that Aronoff (1994:83) criticizes the use of rules of referral, in certain analyses. It is thus worth demonstrating that there are instances of syncretism that are definitely not symmetrical. The Russian example appears well-founded. However, Slavic provides an even clearer case in Slovene (see Corbett & Fraser 1997 for the data).

Since the possible types of such neutralizations cannot be constrained within the formalism, we should look for constraints to impose externally. This is a part of the general enterprise of Network Morphology. The essential point, however, is that the use of the formal language DATR, which forces us to clarify distinctions often left vague, has led us towards the formulation of a typology.

2.3. Network Morphology: Conclusion

In our first example, gender assignment, we saw how typological work led us to use formal methods to clarify the analysis of the crucial language type. In the second example, syncretism, we saw how the different types of expression in the formal language led us towards a typology. From these cases we conclude that computational linguistics and typology are not opposite poles of linguistics but rather they are complementary approaches.

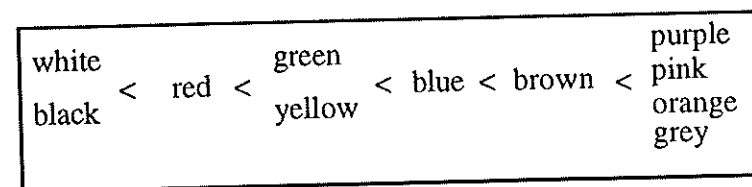
3. Lexical Semantics: the Case of Colour Terms

Our second area is lexical semantics; within this, colour terms form a topic of unique interest, which produces a steady stream of publications by linguists (interested in the implications for linguistic relativity), anthropologists, psychologists, psychophysicists and philosophers (see Hardin & Maffi 1997 for representatives of these different interest groups). For linguists, the Berlin & Kay hypothesis (1969) enjoys a special place, and this confers a unique aura on Russian, as the best-studied problem case.

3.1. Typological Constraints and the Russian System

While the work of Berlin & Kay on basic colour terms continues to provoke interest and research, doubts have remained about their criteria for identifying basic colour terms (as in Cromer 1991:138–140). And so there have been various attempts to find more objective measures, going on alongside extensive empirical work on ever more languages (see Davies, Sosenskaja & Corbett (forthcoming) for a recent example). As originally formulated by Berlin & Kay (1969:5), the hierarchy consists of the following positions:

Figure 3. The Berlin and Kay Hierarchy



The hierarchy constrains the possible inventories of colour terms since the presence of any given term implies the existence of all those to the left (thus a language with a basic term for YELLOW will have basic terms for WHITE, BLACK and RED). It makes diachronic predictions in that languages must move from one state allowed by the hierarchy to another. (Thus a language with basic

terms for WHITE, BLACK, RED and YELLOW would next gain a basic term for GREEN, followed by a basic term for BLUE.) There have been various revisions, in Kay (1975:257-262), Kay & McDaniel (1978:638-640) and Kay, Berlin & Merrifield (1991) and throughout the situation of Russian remains of great interest. It has two basic terms for BLUE, a possibility noted by Berlin & Kay (1969:36, 99) and later in Kay & McDaniel (1978:640-641). Our work has confirmed Russian's unique status with twelve basic terms (instead of the normal maximum of eleven): *belyj* 'white', *čěrnij* 'black', *krasnyj* 'red', *zelěnyj* 'green', *žěltij* 'yellow', *sinij* 'dark blue', *goluboj* 'light blue', *koričnevij* 'brown', *fioletovij* 'purple', *rozovij* 'pink', *oranževij* 'orange', *seryj* 'grey'.⁵

While several of these terms are straightforward, others require comment. Our research suggests very strongly that both terms for blue are indeed basic (see, for example, Corbett & Morgan 1988,⁶ Morgan & Corbett 1989, Davies & Corbett 1994.⁷ Our list varies in two respects from that provided by Slobin for Berlin & Kay (1969:98-99): first we believe the basic term for ORANGE is *oranževij*, and second that for PURPLE is *fioletovij* (see the list experiment in Morgan & Corbett 1989 and the instrumental data on the referent of *fioletovij* in Moss 1989).

⁵ Examples are transliterated in this section.

⁶ Unknown to Corbett and Morgan (1988), Vamling (1986) claimed that Russian has two basic terms for blue, on the basis of frequency in texts. She quoted the list proposed by Kulick and Vamling (1984) which corresponds exactly to the twelve given above, having been established independently. She noted, however, (Vamling 1986: 226) that *fioletovij* 'purple' and *oranževij* 'orange' 'seem to have a less certain status as basic colour terms'.

⁷ For instrumental data on referents of the two terms see Morgan & Moss (1988/89) and for data on children's acquisition of the terms see Davies, Corbett, McGurk & MacDermid (1998). Differences between the terms are treated from the perspective of translation by Alimpieva (1982a) and from a diachronic perspective by Alimpieva (1982b). Examples from early texts are given in Baxilina (1975: 174-207).

3.2. A Formal Approach to Diachrony

It is naturally of interest to consider how such a situation can develop. Let us put that question together with the notion of default inheritance. Taking the historical view, we might claim that by default nothing will happen. Any change may be seen as an override. This provides a means of investigating diachrony in a formal way. We are investigating this idea in joint work involving Ian Davies and Andrew Hippisley, with assistance from Gerald Gazdar, taking the information on the basic colour terms of all the Slavic languages as our starting point (these data are available in Comrie & Corbett 1993). We are attempting to build a computable model of the colour term systems of Slavic, working backwards from current inventories to the earliest times. If we can demonstrate techniques which prove valid where the earlier data are available, they could be employed for other families where the earlier data are lacking. In common with the previous example, this formal approach interfaces with typology, and it also extends into historical linguistics. Initial results are given in Hippisley & Gazdar (1999).

When looking at Slavic, we are looking at a family for which extensive data are available and using a novel method based on the notion of defaults, a method that is computable, so allowing us to check that our claims are valid. We aim to arrive at an account of the colour term systems of the twelve contemporary members of the Slavic language family, of their common ancestor Proto-Slavic, and of the developments which have led to the present situation. We hope to offer a formal, computable approach to diachrony, and a detailed account of the colour system of Slavic, which deserves intensive study given the uniquely exceptional nature of Russian.

There are two main hypotheses we wish to test. First that default inheritance provides feasible reconstructions of ancestor languages, including the common ancestor of the family. Not only will our default inheritance model capture in an elegant way diachronic change in a language family, it will relate these changes in such a way as to reconstruct unchronicled stages in the history of the language family. In the same way, it will be used to give an account of the common ancestor of the language family, in our case

Proto-Slavic. And second that the default inheritance model will reflect the evolutionary dimension of the Berlin and Kay colour term hierarchy. The Berlin and Kay hierarchy in Figure 3 was arrived at by studies on different languages of the world. The claim made is that the universality of the hierarchy is due to the fact that each point on the hierarchy reflects an evolutionary stage in language development. Thus: 'The logical, partial ordering of [the hierarchy]...corresponds, according to our hypothesis, to a temporal-evolutionary ordering...' (Berlin & Kay 1969:4). Note that Proto-Slavic has been claimed to be a Stage V language, that is, having the basic terms as far as BLUE on the hierarchy (Priestly 1981-83:247). This is in itself a hypothesis worth testing. But more important is whether or not our model will yield results consistent with the Berlin and Kay hypothesis according to the most recent modifications. Russian is already problematic for Berlin and Kay, and further inconsistencies with the hierarchy have been found in recent experimental work (Davies & Corbett 1998).

3.3. Lexical Semantics: Conclusion

We hope to demonstrate that a diachronic account can be adequately expressed in DATR, which will allow a degree of rigour and testability not normally available in historical linguistics. This should shed light on an area of special interest in lexical semantics, namely the colour term systems of Slavic, contemporary and historical, and contribute to the typological enterprise initiated by Berlin & Kay.

4. Corpus Linguistics

Our first two formal approaches have both been of the symbolic type. We now turn to one of the stochastic type. Everyone knows that there is a connection between irregularity and frequency (see, for example, Greenberg 1966, Bybee 1995). But there is the question of whether the frequency envisaged is based on the lexeme and all its forms, or just on the irregular form(s). To investigate further we have examined nouns in the Uppsala corpus, a one million word Russian corpus. There are various analytical choices

which had to be made, which are justified at length in the paper on which this section is based (Corbett, Hippisley, Brown & Marriott, forthcoming), and which here we shall take as given. Thus, based on distributional criteria we assume a paradigm of twelve cells, while recognising that no noun has twelve distinct forms; the statistical method too will be accepted without argument here.

The general claim that there is a relationship between frequency and irregularity is something with which almost any linguist would agree. However, that relationship is too vague to be testable. Once we start clarifying the claim, we find an interesting range of possibilities. For instance, we looked initially for a straightforward linear correlation between regularity and frequency; however, the data suggested that it was more appropriate to search for a more complex relationship. Let us start with irregularity and consider its *extent*. Within a given lexeme it might be that every form could be irregular independently; or else it might be that forms come in groups which are regular or irregular together. A second question concerns the *degree* of irregularity. Russian *č'elovek ~ l'ud'-i* 'person ~ people' form an irregular relation, but so do *mést-o ~ mest-á* 'place ~ places'. Intuitively, the first type of irregularity is more severe than the second. If we believe there is a relationship between frequency and irregularity, then we might claim that it will be sensitive to degrees of irregularity. To test this claim we set up a scale of irregularity, devised of course without reference to frequency (section 4.3).

Frequency then can be viewed in two ways. Given a noun whose plural is irregular, with what precisely do we expect to find a relationship? It is easiest to see the alternatives if we consider a corpus and look at the tests we might apply. We might compare lexemes one with another or we could compare regular and irregular forms within lexemes. For the first approach, we could count up how many times each lexeme occurs in the plural. Since we are counting only plurals (without respect to other forms, i.e. the singular) we call this the *absolute frequency* of a lexeme's plural. We can then compare the absolute frequency of plural of different lexemes, to see if there is a relationship between irregular plurals and their absolute frequency. There is, however, a quite different

way to look at the plural (and indeed at any cell or combination of cells in a paradigm). That is to compare it, within the lexeme, with the other available forms. For a given lexeme, we could count how many times it occurs in the plural as compared with the number of times in the singular. This is the *relative frequency* of the plural. We can then compare the relative frequency of the plural in lexemes where it is irregular with that in lexemes where it is regular, as we consider further in the next section.

4.1. Terms and Hypotheses

We now set out a number of hypotheses to test the relationship between irregularity and frequency. We will look for a particular kind of anomaly in the corpus. An anomaly in the plurals of the corpus can be of two distinct types. The first is in terms of an anomalous count of plurals for a lexeme compared to the amount one would expect for a typical lexeme of the corpus; this is *absolute plural anomaly*. What is being compared is an absolute number of plurals for a lexeme with the distribution of the absolute number of plurals in the corpus.

The second type of anomaly is a relative one. Here it is the proportion of instances of the lexeme that are plural which is examined. The distribution of plural proportions can be calculated for the lexemes of the corpus, and if the given lexeme's proportion of plurals is extreme compared to this distribution, we would have identified a *relative plural anomaly*.

We also wish to allow for the possibility of the anomaly being due to a single cell of the paradigm. If one specific cell has an extreme proportion compared to the distribution of the proportion of that cell throughout the corpus, then we have an instance of *cell anomaly*. The anomaly is that a given lexeme has a significantly higher (or lower) than average proportion of word forms for a given cell. We define cell anomaly in relative terms only, because formulating it in absolute terms might lead us to observe plural (or singular) anomaly in disguise. (The cell might be above or below the average simply as a consequence of its singular or plural being above or below the average.)

We will investigate three hypotheses:

- (13) *Hypothesis 1a*
There is a relation between absolute plural anomaly and irregularity

If Hypothesis 1a is confirmed, we will have shown that there is a relation between irregularity and frequency. In order to state the relationship more precisely, we would need to go a little further. If we observed absolute plural anomaly in certain groups of lexemes this might still be because the lexeme as a whole was anomalously frequent. We need a test which will tell us whether the frequency relationship is with the general lexeme, or whether it is specifically with the lexeme's irregular forms. Recall our original question: is frequency related to the lexeme as a whole, or to its irregular word forms? This is provided by Hypothesis 1b.

- (14) *Hypothesis 1b*
There is a relation between relative plural anomaly and irregularity

We will also need to test whether there is a stronger relationship with irregularity when we combine plural anomaly (either absolute or relative) with the more specific cell anomaly. In other words, if a lexeme's plural forms occurred more frequently than average, and a particular cell in the plural was proportionally more frequent than average, are we right in expecting the noun in question to be even more irregular? This is provided by Hypothesis 2, which allows us to look for a stronger (and more fine grained) relationship with irregularity.

- (15) *Hypothesis 2*
Given Hypothesis 1a or Hypothesis 1b is true, there is a stronger relationship between irregularity and the combination of plural anomaly and cell anomaly

A particular case and number may occur more frequently than average either due to the lexeme occurring frequently or to the fact that the cell occurs unusually out of proportion to all word forms in the corpus (absolute frequency of the cell).

4.2. The Data

We test the hypotheses on the nouns in a corpus. Russian is a good choice for this type of investigation. First noun paradigms have sufficient cells for us to tease apart the irregularity of the lexeme in a sub-paradigm, and that of one of its word-forms. Second, irregularity in Russian is highly varied, ranging from full suppletion to shift in stress. We use the Uppsala corpus, which is a set of Russian sub-corpora of various genres, containing in total about one million words. It is considered the best Russian corpus available, in terms of scope and design. For information on the corpus, see Lönngren (1993). The dataset which we created is in the form of a Microsoft Excel document.⁸ Since we were interested in estimating proportions in different categories, there would be large standard errors in our estimates where observed numbers in each category are small. Large sampling errors would complicate detailed cluster analysis. For this reason we recorded only those lexemes which occur at least five times. Our dataset contains around 5440 lexemes, accounting for around 243 000 word forms from the entire one million word corpus.

4.3. The Irregularity Scale

We specifically wish to tease apart the irregularity of a lexeme and that of one of its inflectional forms. We expect a regular noun to have a single (unchanging) stem, as part of that, a fixed stress, and a consistent set of endings. We treat each irregularity type as a step away from regularity. Suppletion is the most severe type of irregularity but even this does not define an end point, since a noun

⁸ The basic dataset is available on the world wide web, and can be found at: <http://surrey.ac.uk/LIS/SMG>, along with a readme file.

with suppletive stems and irregular inflections is more irregular than a noun with suppletive stems but regular inflections. We are investigating *structural irregularity*, i.e. irregularity determined by comparing forms according to a set of principles. Since we wish to investigate the relationship with frequency, we must exclude any frequency consideration when determining regularity. For determining paradigms, we start from the distributional criterion, that is, we determine how many distinctions are justified by the syntax (Comrie 1986, 1991). We accept the traditional view of six cases and two numbers, hence twelve cells in all. We propose the following scale, which is justified in Corbett, Hippisley, Brown & Marriott (forthcoming).

(16) Irregularity Scale

suppletion >
 pluralia tantum >
 stem augments >
 segmental stem irregularity >
 stress stem irregularity >
 segmental inflectional irregularity >
 stress inflectional irregularity >
 full regularity

4.4. Discussion of Results

Our results proved to be extremely interesting. We found relations between frequency and irregularity and a certain degree of correspondence with the Irregularity Scale. We also found evidence for a split between prosodic and non-prosodic morphology.

Absolute Plural Anomaly. The first of our hypotheses, Hypothesis 1a, was confirmed. There is a relation between absolute plural anomaly and irregularity. Below we give eight groups of nouns from the corpus divided up according to our Irregularity Scale; we made a further distinction between two stress patterns which divide the singular and plural. These patterns are, according to the classification in Zaliznjak (1977): pattern C (stem stress

throughout singular, ending stress throughout plural); pattern D (ending stress throughout singular, stem stress throughout plural). The eight groups are given in Table 2.

Table 2. Absolute Plural Anomaly in Eight Groups of Nouns

Group	Type of irregularity	Stress Pattern	Median plural count	Observed number of types	p-value ⁹
1	end stress pl	C	9	64	< 0.001
2	end stress sg	D	5	80	< 0.05
3	stem stress alternation	n/a	22	2	0.25
4	stem alternation	n/a	96	3	< 0.001
5	stem aug in pl	n/a	10	24	< 0.001
6	stem aug in sg	n/a	15	10	< 0.05
7	stem aug in both	n/a	14	14	< 0.05
8	suppletion	n/a	935.5	3	< 0.001

For each of the groups in Table 2 the median value for plural occurrences was significantly higher, as the p-values show, than for the corpus as a whole, with the single exception of Group 3. If we list the groups according to the median value, we get the following: Group 2, Group 1, Group 5, Group 7, Group 6, Group 3, Group 4, Group 8. The data do not support irrefutably the ordering given here as the differences in some cases are insignificant. We cannot reject an ordering of the groups according to the indexing we gave them in Section 4.3. In fact the data here could still be consistent with the principled ordering of the Irregularity Scale, which is an interesting result. Groups 3 and 4 have small sample sizes and their place in the ordering may well be anomalous.

What is shown definitely is that both singular augments and plural augments are related to absolute plural anomaly. While we might argue that singular augments mark the unexpected number

⁹ The p-value represents the probability that a median value more extreme than that observed could have occurred purely by chance. A value < 0.05 is reasonable evidence that there is a relationship between anomaly and irregularity. A value < 0.01 is strong evidence that there is a relationship.

with plural anomaly, this cannot be the case with plural augments, which mark what is the expected number. In other words, it appears that having an augment throughout a particular number (irrespective of whether it is singular or plural) is related to a lexeme having a high plural anomaly. We might have expected an augment in the plural to be associated with higher occurrence of singulars than the average for the corpus. The opposite is the case. In sum, there is a relationship between frequency and irregularity in absolute terms. We must now test our Hypothesis 1b in order to see if this is true in relative terms.

Relative Plural Anomaly. The groups 1-8 were tested for the next of our hypotheses. Evidence for Hypothesis 1b turned out to be not as strong as that for Hypothesis 1a, and involved groups of a specific type. We found evidence for Hypothesis 1b for two groups, and arguably for a third. The stronger evidence is for group 6 (where there is a stem augment in the singular), and group 5 (where there is a stem augment in the plural), and the weaker evidence is for group 4 (where there is a stem alternation). In each case the irregularity is segmental rather than prosodic. The results are given in Table 3.

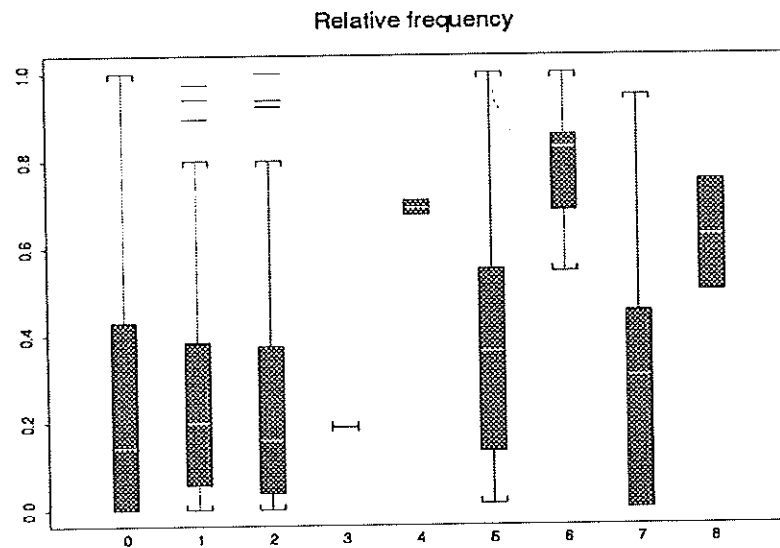
Table 3. Relative Plural Anomaly

Group	Type of irregularity	Median plural proportion	p-value
1	end stress pl	0.2	0.1
2	end stress sg	0.15	0.54
3	stem stress alternation	0.18	0.54
4	stem alternation	0.68	0.06
5	stem aug in pl	0.36	0.03
6	stem aug in sg	0.82	< 0.001
7	stem aug in both	0.32	0.4
8	suppletion	0.62	0.16

Thus we find some evidence that the frequency of occurrence of the irregular forms, and not just frequency of occurrence of the lexeme as a whole, does relate to irregularity of the forms in question. However, frequency of occurrence of forms which are irregular

only in terms of stress does not appear to relate to irregularity. In the box plot in Figure 4 the prosodic groups (Groups 1, 2 and 3) have much lower medians than the others.¹⁰ The median is represented by the white line in the middle of the box; the box itself represents a range of proportions covering the middle 50% of the lexemes in the category; the whiskers cover the remaining 50%, except outliers which are indicated separately with horizontal bars (Daley, Hand, Jones, Lunn & McConway 1995).

Figure 4. Irregularity Type and Plural Anomaly



Key: y axis = proportion of plurals, x axis = irregularity type: 0 = regular, 1 = stress C, 2 = stress D, 3 = stem stress alternation, 4 = stem segment alternation, 5 = stem augment in plural, 6 = stem augment in singular, 7 = different stem augment in singular and plural, 8 = suppletion

¹⁰ Recall that for Hypothesis 1a this does not exclude a relationship between absolute frequency and irregularity for these prosodic irregularities.

It is an extremely interesting and significant result to find that relative frequency of occurrence in the plural appears to be important where non-prosodic irregularity is concerned, but not where prosodic irregularity is concerned. Thus degree of irregularity matters.

Cell Anomaly. Delving deeper into the paradigm, we looked to see if frequency of occurrence of individual case and number cells could be related to their irregularity. We looked at the absolute frequency of occurrences for all cells of given lexemes with one individual irregular cell.¹¹ This is in order to address Hypothesis 2, which is looking for a stronger relationship based on cell irregularity and cell anomaly. Since we are looking for an effect not caused by Hypothesis 1, we must concentrate on cells which do not have a significantly high lexeme frequency. Having investigated this (the nominative plural proved the best candidate) we found little evidence for Hypothesis 2.

4.5. Frequency and Irregularity: Conclusions

Our Hypothesis 1a, that there is a relation between absolute plural anomaly and irregularity, is strongly confirmed. More specifically, nouns which have an irregularity involving a split between singular and plural will tend to be nouns which occur frequently in the plural. There is a less dramatic but still significant effect when only stress is involved. There are some indications of a relation between the degree of irregularity, and the degree of plural anomaly, with cases of suppletion being an extreme case.

Hypothesis 1b, that there is a relation between relative plural anomaly and irregularity was less strongly confirmed. Here we are concerned with the plural forms of a lexeme as a proportion of all its occurrences. Where we did observe an effect, where the plural was used in proportion to the singular significantly more frequently than

¹¹ This includes lexemes for which the cell in question is the only irregularity, as well as lexemes for which the cell irregularity is accompanied by a singular-plural irregularity defined independently of that cell irregularity.

found generally through the corpus, the irregularity was always a segmental one. Furthermore whether the irregularity concerns the singular or the plural, we still find a high relative plural frequency.

When we moved down to examine single cells (Hypothesis 2), we found no evidence that irregularity is related to a high relative frequency of a specific cell in the paradigm, once the effects discussed under Hypothesis 1a and 1b are factored out. This is an interesting result, since it implies a structuring of lexical items. It suggests that an individual irregular cell does not stand out from its subparadigm (singular or plural) in terms of frequency.

There is a relation between frequency and irregularity but this claim is so general as to be relatively uninteresting. Once we clarify the claim, using a formal approach, we see that the relation is more intricate and interesting than we imagined. We find the strongest relation in the "middle ground" where we consider lexemes by splitting them into singular and plural sub-paradigms.

5. General Conclusion

Following the impetus of the FASL series, we have looked at a variety of *Formal Approaches to Slavic Linguistics*. Each one highlights an area of special interest within Slavic and each interfaces with different branches of linguistics, notably typology.

References

- Alimpieva, R. V. 1982a. "Slova *sinij, goluboj* v proizvedenijax S. Esenina i ix ěkvivalenty v pol'skix perevodax". *Materialy po rusko-slavjanskomu jazykoznaniju*, 8-14. Voronež.
- Alimpieva, R. V. 1982b. "Stanovlenie leksiko-semantičeskix grupp cvetovyx prilagatel'nyx v ruskom jazyke pervoj poloviny XIX v.". *Voprosy semantiki: Issledovanija po istoričeskoj semantike*, 49-60. Kaliningrad.
- Aronoff, M. 1994. *Morphology By Itself: Stems and Inflectional Classes* (=Linguistic Inquiry Monograph 22). Cambridge, MA.: MIT Press.

- Baxilina, N. 1975. *Istorija cvetooboznačenij v ruskom jazyke*. Moskva: Nauka.
- Berlin, B. & Kay, P. 1969. *Basic Colour Terms: Their Universality and Evolution*. Berkley: University of California Press. [Revised edition 1991]
- Brown, D. 1998. From the General to the Exceptional: a Network Morphology account of Russian nominal inflection. PhD Dissertation, University of Surrey.
- Brown, D. P., Corbett, G. G., Fraser, N. M., Hippiisley, A. & Timberlake, A. 1996. "Russian Noun Stress and Network Morphology". *Linguistics* 34, 53-107.
- Brown, D. & Hippiisley, A. 1994. "Conflict in Russian Genitive Plural Assignment: a Solution Represented in DATR". *Journal of Slavic Linguistics* 2, 30-48.
- Bybee, J. 1995. "Regular Morphology and the Lexicon". *Language and Cognitive Processes* 10, 425-455.
- Comrie, B. 1986. "On Delimiting Cases". In R. D. Brecht & J. S. Levine (eds.) *Case in Slavic*, 86-106. Columbus, Ohio: Slavica.
- Comrie, B. 1991. "Form and Function in Identifying Cases". In F. Plank (ed.) *Paradigms: The Economy of Inflection* (=Empirical Approaches to Language Typology 9), 41-55. Berlin: Mouton de Gruyter.
- Comrie, B. & Corbett, G. G. (eds.) 1993. *The Slavonic Languages*. London: Routledge.
- Corbett, G. G. 1982. "Gender in Russian: an Account of Gender Specification and its Relationship to Declension". *Russian Linguistics* 6, 197-232.
- Corbett, G. G. 1991. *Gender*. Cambridge: Cambridge University Press.
- Corbett, G. G. & Davies, I. R. L. 1997. "Establishing Basic Colour Terms: Measures and Techniques". In C. L. Hardin & L. Maffi (eds.) *Color Categories in Thought and Language*, 197-223. Cambridge: Cambridge University Press.
- Corbett, G. G. & Fraser, N. M. 1993. "Network Morphology: A DATR Account of Russian Nominal Inflection". *Journal of Linguistics* 29, 113-42.

- Corbett, G. G. & Fraser, N. M. 1997. "Vyčislitel'naja lingvistika i tipologija". *Vestnik MGU: Serija 9: Filologija* no. 2, 122-140.
- Corbett, G. G., Hippisley, A., Brown, D. & Marriott, P. forthcoming. "Frequency and Regularity Revisited: Knowing what to Count". Paper presented at the Symposium "Frequency Effects and Emergent Grammar", Carnegie-Mellon University, 28-30 May 1999, and to appear in the Proceedings, to be published by John Benjamins.
- Corbett, G. G. & Morgan, G. 1988. "Colour Terms in Russian: Reflections of Typological Constraints in a Single Language". *Journal of Linguistics* 24, 31-64.
- Cromer, R. F. 1991. *Language and Thought in Normal and Handicapped Children*. Oxford: Blackwell.
- Daelemans, W., De Smedt, K., & Gazdar, G. 1992. "Inheritance in Natural Language Processing". *Computational Linguistics* 18, 205-218.
- Daley, F., Hand, D., Jones, C., Lunn, D. & McConway, K. 1995. *Elements of Statistics*. London: Addison-Wesley.
- Davies, I. R. L. & Corbett, G. G. 1994. "The Basic Colour Terms of Russian". *Linguistics* 32, 65-89.
- Davies, I. R. L. & Corbett, G. G. 1998. "A Cross-cultural Study of Colour-groupings: Tests of the Perceptual-physiology Account of Colour Universals". *Ethos* 26, 338-360.
- Davies, I. R. L., Corbett, G. G., McGurk, H. & MacDermid, C. 1998. "A Developmental Study of the Acquisition of Russian Colour Terms". *Journal of Child Language* 25, 395-417.
- Davies, I. R. L., Sosenskaja, T. & Corbett, G. G. Forthcoming. "First Account of the Basic Colour Terms of a Daghestanian Language: the Case of Tsakhur". To appear in: *Journal of Linguistic Typology*.
- Evans, R. & Gazdar, G. 1989a. "Inference in DATR". *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics*, 66-71. Manchester, England.

- Evans, R. & Gazdar, G. 1989b. "The Semantics of DATR". In A. G. Cohn (ed.) *Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 79-87. London: Pitman/Morgan Kaufmann.
- Evans, R. & Gazdar, G. 1996. "DATR: a Language for Lexical Knowledge Representation". *Computational Linguistics* 22, 167-216.
- Fraser, N. M. & Corbett, G. G. 1995. "Gender, Animacy, and Declensional Class Assignment: a Unified Account for Russian". In G. Booij & J. van Marle (eds.) *Yearbook of Morphology 1994*, 123-50. Dordrecht: Kluwer.
- Fraser, N. M. & Corbett, G. G. 1997. "Defaults in Arapesh". *Lingua* 103, 25-57.
- Gazdar, G. 1987. "Linguistic Applications of Default Inheritance Mechanisms". In P. Whitelock, M. McGee Wood, H. L. Somers, R. L. Johnson & P. Bennett (eds.) *Linguistic Theory and Computer Applications*, 37-68. London: Academic Press.
- Gazdar, G. 1990. "An Introduction to DATR". In R. Evans & G. Gazdar (eds.) *The DATR Papers*. Cognitive Science Research Paper CSRP 139, 1-14. School of Cognitive and Computing Sciences, University of Sussex.
- Gazdar, G. forthcoming. "Ceteris Paribus". To appear in J. A. W. Kamp & C. Rohrer (eds.) *Aspects of Computational Linguistics*. Berlin: Springer.
- Greenberg, J. 1966. *Language Universals, with Special Reference to Feature Hierarchies*. The Hague: Mouton.
- Hardin, C. L. & Maffi, L. (eds.) 1997. *Color Categories in Thought and Language*. Cambridge: Cambridge University Press.
- Hippisley, A. 1997. Declarative Derivation. Ph.D. Dissertation, University of Surrey.
- Hippisley, A. & Gazdar, G. 1999. "Inheritance Hierarchies and Historical Reconstruction: towards a History of Slavonic Colour Terms". Paper read at the 35th Regional Meeting of the Chicago Linguistic Society, April 1999.

- Huntley, D. 1993. "Old Church Slavonic". In B. Comrie & G. G. Corbett (eds.), *The Slavonic Languages*, 125-187. London: Routledge.
- Kay, P. 1975. "Synchronic Variability and Diachronic Change in Basic Color Terms". *Language in Society* 4, 257-70.
- Kay, P. & McDaniel, C. 1978. "The Linguistic Significance of the Meanings of Basic Colour Terms". *Language* 54, 3, 610-646.
- Kay, P., Berlin, B. & Merrifield, W. 1991. "Biocultural Implications of Systems of Color Naming". *Journal of Linguistic Anthropology* 1, 12-25.
- Kulick, Don & Vamling, Karina 1984. "Ryska". In T. Pettersson (ed.) *Färgterminologi: Seminarieuppsatser i Allmän Språkvetenskap* (=Praktisk Lingvistik 9), 79-109. Lund. [Quoted from Vamling (1986).]
- Lönngren, L. 1993. *Častotnyj slovar' sovremennogo russkogo jazyka*. (=Acta Universitatis Upsaliensis, Studia Slavica Upsaliensis 33). Uppsala.
- Morgan, G. & Corbett, G. G. 1989. "Russian Colour Term Saliency". *Russian Linguistics* 13, 125-141.
- Morgan, G. & Moss, A. E. St. G. 1988/89. "The two Blues of Russian: the Referents of *sinij* and *goluboj*". *Die Farbe* 35/36, 353-357.
- Moss, A. E. 1989. "Basic Colour Terms: Problems and Hypotheses". *Lingua* 78, 313-320.
- Perlmutter, D. M. & Orešnik, J. 1973. "Language-particular Rules and Explanation in Syntax". In S. R. Anderson & P. Kiparsky (eds.), *A Festschrift for Morris Halle*, 419-59. New York: Holt Rinehart.
- Priestly, T. M. S. 1981-83. "On Basic Color Terms in Early Slavic and Ukrainian". *The Annals of the Ukrainian Academy of Arts and Sciences in the United States*, vol. XV [appeared 1987], nos 39-40, 243-251.
- Priestly, T. M. S. 1993. "Slovene". In B. Comrie & G. G. Corbett.(eds.), *The Slavonic Languages*, 388-451. London: Routledge.

- Shieber, S. M. 1987. "Separating Linguistic Analyses from Linguistic Theories". In P. Whitelock, M. McGee Wood, H. L. Somers, R. L. Johnson & P. Bennett (eds.) *Linguistic Theory and Computer Applications*, 1-36. London: Academic Press.
- Spencer, A. 1991. *Morphological theory*. Oxford: Blackwell.
- Vamling, Karina 1986. "A Note on Russian blues". *Slavica Lundensia* 10, 225-33.
- Zaliznjak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka: slovoizmenenie*. Moscow: Russkij jazyk
- Zwicky, A. 1985. "How to Describe Inflection". In M. Niepokuj, M. Van Clay, V. Nikiforidou & D. Feder (eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, 372-386. Berkeley, California: B. L. S., University of California.

SLIS, University of Surrey,
Guildford, Surrey, GU2 5XH, GB
g.corbett@surrey.ac.uk