

## Chapter 2

### Features: essential notions<sup>1</sup>

Greville G. Corbett

Preprint July 2010:

To appear in: Anna Kibort and Greville G. Corbett (eds)

*Features: perspectives on a key notion in linguistics.*

*Oxford: Oxford University Press*

### 2.1 Introduction

Features have been central to linguistics, implicitly or explicitly, from the earliest times. They are ‘standard currency’: in particular, each of the various syntactic frameworks relies heavily on them. To understand why, one has only to attempt a description of English syntax without features. Rule after rule, or constraint after constraint, would have to be duplicated to allow for singular and plural forms (see Gazdar and Mellish 1989: 218-219 for a comparable demonstration from French). Features thus allow generalizations in syntax; they do so similarly in morphology (Corbett and Baerman 2006). And once we deal with orthogonal features, say case and number, the savings are substantial. Computational work too uses features in numerous applications (see Copestake this volume). While features are heavily used, they are often taken for granted. In fact the level of confusion (much of it not recognized) is considerable. This is a missed opportunity, since features are where different approaches converge. That is why we have put together this volume.

This chapter discusses basic issues, attempting to unpick some of the assumptions commonly made about features. And it includes some of the ways in which the field is moving forward. I hope that as a result linguists will make more conscious choices in the use of features.

## **2.2 The use of features**

As the use of features has expanded and developed, various distinctions have been drawn, and different conventions have arisen. It is worth revisiting these, to check which we adopt for good reason and which may have become no more than unjustified habits.

### *2.2.1 Features for different components*

Features are partial descriptions of linguistic objects; as such they allow us to capture regularities in different components. We therefore recognize features which apply within a given component, for example, semantic, syntactic, morphological and phonological features.

There are also features which have an effect across component boundaries. Perhaps the best known are the morphosyntactic features (Matthews 1972: 162). On a strict definition, any such feature must have a role in both components, that is, morphosyntactic features must have role in both syntax and morphology, they are not just the sum of morphological and syntactic features. Such features may be termed ‘interface features’ (Svenonius 2007). It is worth sharpening our definitions, to distinguish morphosemantic features (with no role in syntax) from morphosyntactic features (which evidently have a role in syntax; compare Stump 2005: 52). Features like tense and aspect are often morphosemantic rather than strictly morphosyntactic.

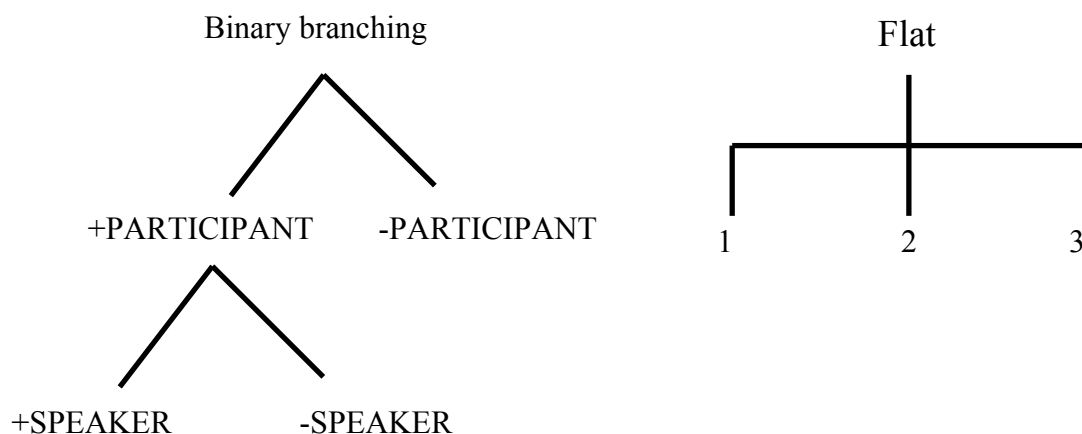
Provided we can define and distinguish the feature types clearly, we can maintain certain interesting claims. First, syntax is phonology-free (Pullum and Zwicky 1988); that is, syntactic rules cannot refer to phonological features. A rule of the type ‘vowel-initial verbs are clause-final’ is excluded. And second, syntax is also morphology-free (Zwicky 1996: 301, Corbett and Baerman 2006); syntactic rules cannot refer to morphological features. A rule of the type ‘verbs of inflectional class II are clause-final’ is excluded.

### *2.2.2 The internal structure of features*

Some linguists assume that features are binary, while others allow for larger numbers of values.<sup>2</sup> Halle (1957) gives arguments for adopting binary features in phonetics and phonology, but this is unusual; the internal structure of features tends to be assumed rather than argued for. Assuming binarity leads some researchers to decompose feature values. Jakobson (1958) represents a heroic failure in this regard, and there have been some less heroic campaigns since.<sup>3</sup>

When features appear to have more than two values, they can of course be split and be represented by additional binary features. This may mean there are superfluous values; this can be seen if three possibilities are simply represented by two binary features, and the issue is more serious with features with awkward numbers of values like five or nine. Alternatively, a geometry may be proposed in addition, making one feature subordinate to the other. Let us take a simple instance, a three-person system. An account with hierarchical binary branching is given below, along with the flat structure.

Fig. N.1 Feature structures: person



It is often claimed that syncretism provides evidence to support binary branching structures. The branching structure allows of course for all person forms to be different, but also for first and second persons to be syncretic (by reference to the +PARTICIPANT node), or for all three to be syncretic (by reference to the root node). Now there are indeed languages which show syncretism of first and second persons, in the non-singular; these include Burarra, Dogon, and Manchad (Baerman, Brown and Corbett 2005: 59). But now consider syncretisms like that found in German, where first plural and third plural are identical for all verbs (*wir finden* ‘we find, *sie finden* ‘they find’).<sup>4</sup> Here we see that the additional structure in the binary branching version is of no value in capturing the syncretism. We need to appeal to a different mechanism. We could specify that the –SPEAKER form is *findet*, and that by default the remaining plural forms are *finden*.

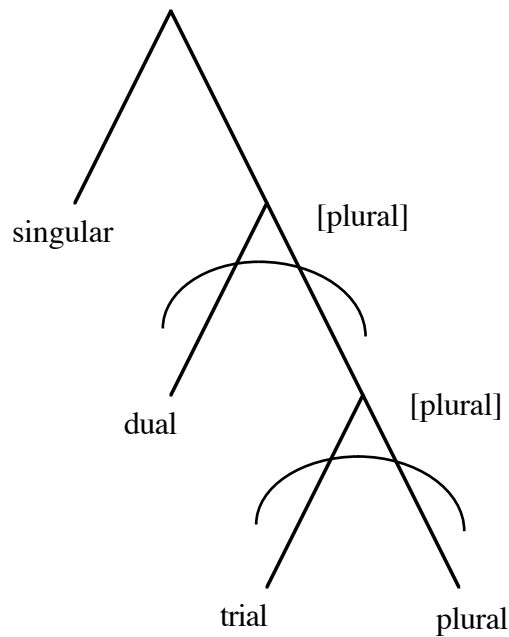
Now let us compare the flat structure. We can treat the syncretism of first and second persons, by specifying the third person and having a default form for the first and second persons. For the German example, we can specify the second person form as *findet*, and a default form *finden* (for the remaining two unspecified forms). Thus the flat structure allows us to capture both types of syncretism. The binary branching

structure gives no advantage: in the German example the structure is actually superfluous. This means that with small systems like this, the morphology actually provides no argument for a particular structure. The picture becomes more complex and interesting with larger numbers of values; see Baerman, Brown and Corbett (2005: 59-61, 126-133) for illustration and analysis. It is shown there that attested features with larger numbers of values pose different problems for the two approaches.

My point here is the more basic one, namely that the internal structure of features is something to be argued for rather than merely assumed. Having pointed out an instance where we should have considerable reservations about a hierarchical feature structure, I should stress that equally there are instances where there *are* good reasons for postulating internally structured features. Evidence comes from the related issues of facultative values and superclassing.

We talk of **facultative values** when a feature has one or more values which the speaker is not required to use. Thus in Larike (a Central Moluccan language spoken on the western tip of Ambon Island, Indonesia) the number feature has the values singular, plural, dual and trial; the trial is strictly for three referents. However, the trial need not be used for three referents, nor indeed the dual for two. These values are facultative: in their place the plural may be used (Laidig and Laidig 1990: 93, Corbett 2000: 44-45). The number feature as a whole is not optional: the singular is not used in place of the plural. Here the value which is employed when the facultative value is not chosen, that is, plural for trial and plural for dual, reveals the structuring of the feature values, as represented in figure 2.2. The arcs indicate the facultative portions of the structure.

Fig. 2.2 Facultative number values in Larike



**Superclassing** is a more restricted phenomenon, in that here the choice of use is restricted to agreement. In superclassing, some but not all of the available distinctions are drawn; see the account of Bininj Gun-wok (Mayali) in Evans (1997: 127-140).

Superclassing is also found in Jingulu, a non-Pama-Nyungan language of the Northern Territory of Australia (Pensalfini 2003 and personal communication). Like Bininj Gun-Wok, Jingulu has four genders: masculine, feminine, vegetable and neuter. Gender assignment is largely a matter of semantics: nouns denoting male animates are masculine, those denoting female animates are feminine, edible plants are vegetable and the residue neuter. However, there are additional principles and some instances where the gender value of a given noun is hard to understand.

Adjectives agree in gender as follows:

Jingulu (Pensalfini 2003: 160-161, 164-167, Corbett 2006: 151-154)

(1) Lalija darra-nga-ju **jamurriyak-a**.

tea(M) eat-1SG-do cooled-M

‘I’m drinking cold tea.’

(2) Wijbirri-rni **jalyamingk-irni**.

white.person-F new-F

‘The white girl is new-born.’

(3) Miringmi-rni darra-nga-yi **bardakurr-imi**.

gum(VEG)-FOC eat-1SG-FUT good-VEG

‘I’ll eat the sweet gum.’

(4) Jami-rna dimana-rni laja-ardu **ngamulu** lanbu.

that.M-FOC horse(M)-ERG<sup>5</sup> carry-go big.N load(N)

‘That horse is carrying a big load.’

The examples we have seen show full agreement, demonstrating the existence of four genders. However, sometimes we find less than full agreement, as in these examples:

(5) Ngamulirni **jalyamungk-a** binjiya-ju, birnmirrini.

girl(F) young-M grow-do prepubescent.girl

‘That little girl is growing up into a big girl.’

These next were offered by a speaker as alternatives:

(6) **ngininiki** barndumi *or* **ngimaniki** barndumi

this.N lower.back(VEG) this.VEG lower.back(VEG)

‘this lower back’

‘this lower back’

These two examples show superclassing, with masculine for feminine in (5), and neuter for vegetable in (6). The choices are not random: these are the possibilities for superclassing. The masculine can also be used as the ultimate default:

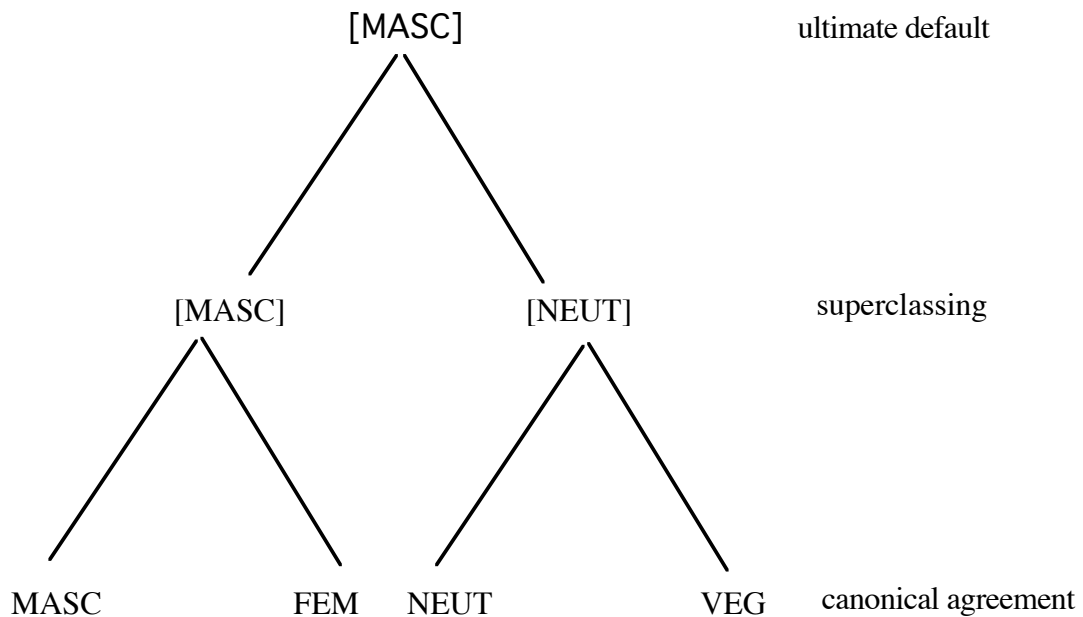
(7) **Jama**-rni nyanyalu-ngkujku, darrangku kirdkilyaku.

that.M-FOC leaf-HAVING.N tree(N) bent.N

‘That bent tree is leafy.’

We may represent this situation as in figure 2.3:

Fig. 2.3 Gender superclassing in Jingula



At the lowest level we have full (canonical) agreement. Then at a level up we have superclassing, where only some of the potential distinctions are drawn. And at the top level we have the ultimate default, which is the masculine in this system. Thus we have not only a default value, but also a sort of half-way-house, which we call ‘superclassing’. The point is that we have an optional collapsing of forms. Since the masculine is the ultimate default, the argument depends on the possibility of neuter and of masculine agreement for vegetable nouns. In the case of feminines, there is no way to distinguish the masculine which might be expected for superclassing from the masculine as the ultimate default.

Looked at abstractly enough, the situation could be described as a kind of syncretism. But it differs significantly from the standard examples in respect of the

context. For standard syncretism we specify contexts in morphological terms or in morphosyntactic terms; for example, the nominative is syncretic with the accusative in inflection class II (a morphological context), or in the plural (a morphosyntactic context). In superclassing the speaker has choices available, unrestricted by grammatical considerations. The choices are: the unique realization for the featural specification and a general default (this much is a common situation), and the third possibility, that is the speaker can mark agreement but not full agreement, reflecting the superclass of the appropriate feature value. And these simultaneous options do not depend on morphological or morphosyntactic context. Thus superclassing represents options available to the speaker, and the structure of those options can be seen as evidence that the values of the feature are structured.

### *2.2.3 Feature values are not equal*

Let us now consider instances where, unlike those just discussed, there is no such obvious evidence for structuring. This would appear to be so when a feature has only two values. Yet we still find that the values are not equal. We can see this in at least four ways.

First we can simply look at frequency (see, for instance, Corbett 2000: 280-281 for a compilation of data). In various Indo-European languages, the sources reported always find the singular used more frequently than the plural (typically the singular is used in around three quarters of the instances).

Second, we look at items which lack a value. We typically find that these exhibit non-random patterns. Thus in English, there are many nouns which do not have both number values. Many abstract nouns like *health*, *wealth* and *happiness* have no plural; however, many can have a plural when recategorized (*a particular happiness* and *particular happinesses*). Those with no singular, like *scissors*,

*binoculars* and *trousers*, typically denote concrete objects in English. They have no plural; they have half an explanation, in that they typically denote paired objects, yet not all paired objects (like *bicycle*, *bigraph*, *dromedary*) behave in this way. The patterns vary across languages, but in general the lack of particular values is not random.

Third we should consider examples like this:

(8) To err is human.

One analysis of such constructions is that the verb has to agree, and yet there is no controller with the required feature specification. In such circumstances there is a ‘least worst’ option, which in English is the third person and singular number. Some might wish to attribute the choice to markedness considerations; this is a problematic step (see Haspelmath 2006 for the difficulties with markedness). It is at least worth pointing out that the choice of the feature value differs across languages, sometimes for number (Corbett 2000: 185-186) and often for gender (Corbett 1991: 203-212).

And fourth, we should note the use of ‘evasive’ feature values. Generally, when a choice between feature values is problematic, one of them is chosen. Thus if there is a choice between masculine and feminine, and the speaker does not know which (for instance, in asking a question), the particular language specifies masculine or feminine. Occasionally a third value is chosen (thus given an awkward choice between A and B, the form used is C). This is an ‘evasive’ use. For instance, in the Daghestanian language Archi, there are four gender values: gender I for male humans and gender II for female humans; gender III includes most other animates, and some inanimates, and IV has the remaining nouns. The word *lo* ‘child’ can be used with agreements for I (male human) or II (female human), as we might expect. In addition, the use of gender IV agreements is also possible (this is the gender for abstracts, some

inanimates, and rather few animates). This use ‘evades’ the two obvious genders and so makes no commitment as to the sex of the child. Again the choice of the evasive feature value varies across languages.

Such inequalities are modelled in different ways, notably through the use of defaults. While linguists have rightly embraced the notion of defaults with enthusiasm, we should note that the term is used in rather different senses. We say that by default English verbs form the past tense in *-ed*. This default is then overruled for various exceptional verbs. This is a ‘normal case default’, according to Fraser and Corbett (1997: 44). Such instances are rather different from the type of default which comes into play to cover various different non-canonical situations (for instance, agreement with controllers which lack the relevant specification, including those which are totally absent, as in impersonals). This type is termed an ‘emergency case default’ (Fraser and Corbett 1997: 44).

In such instances it is important to be alert to the possibility of smuggling in additional feature values (Stanley 1967: 409-411 is an early warning with respect to this issue). If one is careless, the feature number, for instance, can have the values singular, plural and unspecified, while apparently being a binary feature. This is a something to be alert about when notions like ‘valued’ and ‘unvalued’ are introduced.

#### *2.2.4 Features are not equal*

Having noted that feature values are not equal, we should also recognize that features themselves are not equal, irrespective of their values. Consider a simple example like *stars shine*. Both subject and predicate mark number, but the feature is rather different for each. *Stars* is plural for good semantic reasons, and if we change number to the singular value, that would equally represent a semantic choice. On the other hand, the plural of *shine* is not motivated semantically. There is not necessarily more than one



One solution is to suggest that possessive adjectives of this type have both inherent and contextual features of number and gender, and that their values are independent of each other (see Stump 2001: 15-17). In example (9), *mužowa* is contextually feminine and singular (agreement with *sotra*) and is inherently masculine and singular (controlling *mojeho*). Possessive adjectives of this type in Sorbian are always inherently singular; they may be masculine or feminine. Contextually they may take any number (singular, dual or plural). Case is less clear. I have treated the case of the phrase as being externally determined (rather than as a matter of agreement within the phrase). Either way, the case of *mužowa* and *sotra* are contextually determined. But then to account for the case of *mojeho*, we need to say that *mužowa* governs the genitive (which is plausible at least in part, given that it is formed with a possessive suffix).

The key point, then, is that features may be inherent or contextual, and that the same feature may be inherent and contextual on one and the same item; the values of the features are then independent of each other. This distinction is the essential component of the distinction between interpretable and non-interpretable features within Minimalism.

### *2.2.5 Constraining feature structures*

It is evident that there are common constraints on feature structures. We find languages like German, which distinguish gender in the singular but not in the plural, but we do not find a hypothetical language German', with gender distinguished in the plural but not in the singular. Constraints of this type were given by Greenberg (1963: 112-113), and are taken up as Feature Cooccurrence Restrictions in GPSG (Gazdar, Klein, Pullum and Sag 1985: 27-29).

More recently, ‘typing’ of feature structures has been introduced. Typing is employed both to state the possible features and appropriate values, and to require that the required values are specified. The key reference is Carpenter (1992); this is a demanding read. Alternatively, Copestake offers ‘a gentle but precise account’ (2002: 3), and Sag, Wasow and Bender (2003: 59-72) is a helpful exposition.<sup>6</sup> This is a clear point of division between HPSG, which uses typing, and LFG, which does not, relying instead on the notions of completeness and coherence.

### *2.2.6 Parts of speech modelled with features*

When features are mentioned with regard to morphology and syntax, the first to come to mind may well be case, number and so on. However, part of speech classification is frequently represented in featural terms, too. There is explicit discussion in Gazdar, Klein, Pullum and Sag (1985: 17-18); they refer to Chomsky (1965) as a predecessor (see 1965: 79-86 and 110-111). From one point of view this is fully consistent: features can perfectly well do the job. Features can also represent the projections of these categories in syntax. And yet it may be significant that when linguists use features like this, a shorthand like NP or DP is often retained too. This preserves the intuition that part of speech categories and subcategories are one type of categorization, and morphosyntactic features are a cross-classification. Thus while it makes sense to use feature notation in both instances, this brings the need to classify the features, according to their rather different functions.

### *2.2.7 The issue of identity*

Identities constitute a considerable part of syntax. We need to guarantee identity of feature values across a range of constructions. The early method of managing this was copying. Feature values were simply copied, from controller to target, and thus they

were bound to be identical. There are various problems with copying, as pointed out by Barlow (1992) and in Pollard and Sag (1994). For instance, take this Russian example:

(10) Russian

Ja                    side-l-a

1SG.NOM          sit-PST-F.SG

‘I was sitting’ (woman talking)

The verb is feminine in this example; if the speaker were male the verb would be masculine. But the form of the subject pronoun remains the same. In a copying account we require two different subject pronouns *ja* ‘I’, which are identical phonologically, simply so that the feature values which are copied can be different.

For this, and other reasons detailed in the sources above, copying is not used in HPSG and LFG. It was retained in GB, and replaced in Minimalism by ‘checking’. This notion lacks a full formalization; for discussion of the issues see Asudeh and Toivonen (2006a: 409-420), Adger (2006) and Asudeh and Toivonen (2006b).

In other approaches, unification has had a major role. This goes back to work by Kay (1979). A valuable entry into this work is the account in Shieber (1986: 14-16). In HPSG, unification is central. However, it is the identity constraints which matter, since ‘unification is but one of the many procedures used to solve systems of identity constraints’ (Ginzburg and Sag 2000: 2, fn. 2). LFG uses equality and satisfiability. Unification is a way of implementing equality; thus unification is a submodule in an implementation (Ron Kaplan, personal communication).<sup>7</sup> In such approaches, our example (10) fits well; the subject has the feature values first person and singular, and these unify with (are compatible with) the feature values of the verb, namely singular and feminine.

While constraint based approaches have had considerable success in using features to handle situations of matching, there are serious problems in those situations where a match might be expected but the feature values do not in fact match. There are two well-established sets of difficult data.

The first concerns lexical hybrids, such as *committee* in different varieties of English (Copestake 1995, Corbett 2006: 211-213). The problem with such hybrids is that their feature specification appears to differ according to the agreement target. For attributive modifiers they are singular (*this committee* is the only possibility, not *\*these committee*). For other targets, both singular and plural may be found.

The choices are not free, but are tightly constrained by the Agreement Hierarchy (Corbett 2006: 207):

(11) The Agreement Hierarchy

attributive > predicate > relative pronoun > personal pronoun

On the basis of this hierarchy, we can constrain possible agreement patterns as follows:

- (12) For any controller that permits alternative agreements, as we move rightwards along the Agreement Hierarchy, the likelihood of agreement with greater semantic justification will increase monotonically.

Staying with *committee* and similar items, here are the results of a large study:

[T/S Table 2.1 here]

The US data are from the Longman Spoken American Corpus (LSAC), which has five million words, and GB data come from the ten million word section of the British National Corpus (BNC) devoted to spoken language.<sup>8</sup> Attributive position is not included, since only singular agreement is found there, as just discussed. The remaining data are clearly in accord with the constraint given in (11).

To demonstrate that the familiar English example is representative of many others, consider the following summary data from a range of languages (details can be found in Corbett 1991: 226-236, 2006: 214-218).

[T/S Table 2.2 here]

Each of the hybrids listed conforms to the constraint of the Agreement Hierarchy, but each is a problem for straightforward unification accounts.

The second set of problems concerns constructional mismatches; these are constructions which have the same properties as lexical hybrids. That is, the agreements they control depend on the nature of the target. Like lexical hybrids, they are subject to the Agreement Hierarchy. The best known example is conjoined noun phrases (which typically allow agreement with just one conjunct, normally the nearest, or with all).<sup>9</sup> Other types of constructional mismatches include pseudo-comitatives, syntactic/semantic head mismatches, and default form versus agreement (Corbett 2006: 208-210, 220-224).

Conjoined noun phrases bring with them the related issue of resolution rules, as illustrated in this classic example:

(13) Slovene (Lenček 1972: 60)

T-o	drev-o	in	gnezd-o	na	njem	mi
that-N.SG	tree(N)-SG	and	nest(N)-SG	on	3SG.N.LOC	1SG.DAT
	<b>bosta</b>		<b>ostal-a</b>	v	spomin-u.	
	AUX.FUT.3DU		remain-M.DU	in	memory-SG.LOC	

‘That tree and the nest on it will remain in my mind.’

The Slovene resolution rules (called into play when agreement is with all conjuncts) determine that two singular conjuncts require a dual target, and two neuters require a masculine target. Hence the masculine dual verb. Recent references on resolution include: Dalrymple and Kaplan (2000), Wechsler and Zlatić (2003: 171-195), Corbett (2006: 238-263).

A further identity problem, which has gone largely unnoticed, is that features values need not match even within a periphrastic construction (Corbett 2006: 86-87). We might have assumed that within a single cell of a paradigm, which is how we may think of a periphrastic form, the values of shared features would be the same. However, this is not necessarily the case:

Czech (Eva Hajicová and Jarmila Panevová p.c.)

- (14) by-l-a            jste            velmi    laskav-á  
         be-PST-F.SG    AUX.2PL    very    kind-F.SG  
         ‘You were very kind’ (addressed to a woman)

Such an expression is appropriate for polite address to a single addressee. No pronoun is included, but *vy* ‘you’ can be included if it is under contrastive stress (*Vy jste byla velmi laskavá*). The auxiliary verb, the clitic *jste*, is second person plural, while the past participle *by-l-a*, literally ‘was’, is singular. We therefore have a periphrastic verb, and its two parts have different values for number.

### 2.2.8 Matching across components

Everyone knows that tense does not match time, though there is a relation between them. In many languages it is obvious that morphosyntactic gender corresponds in part to a semantic distinction, but only in part. However, less evident mismatches across components sometimes pass unnoticed. Thus semantic number and morphosyntactic number are not equivalent; here the variation in the semantic value of number with different lexical items is constrained in a principled way by the Animacy Hierarchy (Corbett 2000: 55-57, 83-87). Similarly a morphological distinction may follow a phonological one, but not fully (thus Russian morphological alternations based on the palatalized versus non-palatalized distinction do not completely follow phonological palatalization). Despite this common knowledge,

some make unwarranted inferences from one component to another, perhaps on occasion misled by the similar names of the features. For instance, there has been some excitement about the apparent structuring of feature values based on the evidence of syncretic patterns; however, Baerman, Brown and Corbett (2005) show that within morphology the possible patterns are extremely varied, and that inferences that morphology directly reflects semantic features and structure are unwise at best.

### **2.3 The inventory of features**

There are several analyses in the literature in which features are sprayed around with disturbing nonchalance. The old warning still applies:

“So linguists fudge, just as has been done in the reflexive rule, by sticking on the arbitrary feature +REFL. Such a feature is a fudge. It might just as well be called +CHOCOLATE, which would in fact be a better name, since it would clearly reveal the nature of the fudge.”

Lakoff (1972: ii)

As a counter to this profligacy, there are aspirations at various points in the literature to a list of features and values. It makes good sense to aim for such a list, which would be a simple typology, unless and until it is proved impossible. Such an inventory requires the solution to two problems, the analysis problem and the correspondence problem.

#### *2.3.1 The analysis problem*

For some languages it is relatively easy to determine the features and their values. In other languages it is a major undertaking; some famous instances have given rise to long-running disputes. A good start on the analysis problem was made by members of the Set-theoretical School, which included scholars such as Kolmogorov, Revzin,

Zaliznjak, and Marcus. A careful and sympathetic survey is provided by van Helden (1993); the review of this work, by Meyer (1994), is a good entry-point into the literature. Given that a detailed account is available in van Helden (1993), a simplified summary will be given here.

Zaliznjak and others worked out careful and consistent methods for determining the feature and value inventory of a language. A first step is to iterate through lexical items and contexts, so long as distinctions are discovered.

[T/S Table 2.3 here]

Where two items produce identical distinctions, the columns can be collapsed. Thus if two items are *cat* and *dog* in English, they will fit identically into contexts such as *this ... (this dog and this cat are equally good)* or *the ... sleeps (the cat sleeps and the dog sleeps are both fine)*. Similarly if two contexts allow identical inventories of items the relevant rows can be collapsed. Consider now a typical instance from Russian, a language where the inflectional morphology has a much greater role than that of English.

[T/S Table 2.4 here]

If we had only the evidence of the first noun *žurnal* ‘magazine’, we would have to say that the contexts 1 and 3 provided no evidence for different values. However, when we put *gazeta* ‘newspaper’ in the same two contexts, this provides evidence for distinct feature values (the traditional nominative and accusative). Quite often the result is that the expected features and values are established, but that less clear instances emerge too. In fact, Russian has arguably ten case values, rather than the traditional six (see Corbett 2008 for discussion).

For each feature and especially for each value, we need rules as to when it is used. These are sometimes termed ‘assignment rules’. I stress that for each feature

and value both justification and assignment rules are required. In work on individual languages one or other may be favoured, according to the difficulty of the issues: thus in one language it may be easy to justify postulating a particular feature, but hard to pin down the rules for its use, while in another, determining the number of values of a feature may be the intellectual challenge which has attracted attention.

### *2.3.2 The correspondence problem*

In analysing and comparing languages we naturally use similar labels for the features proposed. Yet it is not self-evident that a particular feature (say number) corresponds across languages, and the values even less (Saussure 1916/1971: 161, Gazdar cited and discussed in Zwicky 1986a: 988-989). Yet typological work depends on our resolving these issues. We should continue to attempt to prove cross-linguistic validity of our features, through care about definitions, perhaps within a canonical approach (Corbett 2008). We should also note that the problem has an additional twist: even within a single language, feature values do not always correspond straightforwardly across the elements that carry them. This is shown for instance by gender in Romanian (Corbett 1991: 150-151), or number in Bayso (Corbett 2000: 181-183, based on Corbett and Hayward 1987).

## **2.4 Practicalities**

Since features are shared across sub-areas of the discipline, from the highly theoretical to the most applied, there are practical steps which can have general benefit. There are various steps towards standardizing and generalizing. And then we shall refer to instances of using features in large-scale implementations, which serve as a valuable testing-ground for theories of features.

### *2.4.1 Glossing*

Though features are common currency, they can lead to confusion even at the most basic level. Thus we may see PERF used to mean ‘perfect’ in one paper and ‘perfective’ in the next, sometimes without even an account of the abbreviations used. At this level the Leipzig Glossing Rules (Comrie, Haspelmath and Bickel 2004) represent a useful step forward. They give standards for glossing, and propose some standard abbreviations. This is a bare minimum for the discipline.

### *2.4.2 EAGLES (Expert Advisory Group on Language Engineering Standards)*

The report on morphosyntactic annotation of this group (Leech and Wilson 1996) represents an early attempt to grapple with the practical issues raised by features. It was restricted to languages of the European Union, which makes it typologically limited, and it does not fully distinguish parts of speech and semantic subcategories from morphosyntax (see 2.2.6 above). It appears that tags which were suggested for particular languages were included without having been rigorously compared with the general set established for a wider range of languages.

### *2.4.3 The ISO: Lexical markup framework (LMF)*

The International Organization for Standardization (ISO), in particular Technical Committee ISO/TC 37, *Terminology and other language resources*, Subcommittee SC 4, *Language resource management*, worked for several years developing ISO 24613: 2008 ‘Language resource management – Lexical markup framework (LMF)’.

The purpose (from the introduction) is this:

Lexical Markup Framework (LMF) is an abstract metamodel that provides a common, standardized framework for the construction of computational lexicons. LMF ensures the encoding of linguistic

information in a way that enables reusability in different applications and for different tasks. LMF provides a common, shared representation of lexical objects, including morphological, syntactic, and semantic aspects.

LMF was intended for large-scale applications, and seemed a long way from linguists' main interests. Indeed the linguistic content was insecure. The Committee has taken note of some of the concerns of linguists in the published version.

#### *2.4.4 E-MELD (Electronic Metastructure for Endangered Languages Data) and GOLD (General Ontology for Linguistic Description)*

There were two main objectives behind E-MELD: contributing to the preservation of data on endangered languages, and developing the infrastructure for effective collaboration between electronic archives (Aristar Dry 2002). The first objective related to best practice, in various areas. So far as it concerned morphosyntactic markup, the direction was not so much to suggest a standard, as to ensure that non-significant differences in annotation should not hamper further understanding and analysis. This was consonant with the second objective, and led to initial work on an ontology of linguistic concepts (Farrar and Langendoen 2003).

#### *2.4.5 Inventory of the features*

There are references in the literature, usually almost as asides, that there could be an inventory of the features, from which particular languages draw. A list of the features would be the simplest possible typology, and it is surely something we should attempt to achieve. If we discover insurmountable problems we would then reasonably look for more complex alternatives.<sup>10</sup>

#### *2.4.6 Large-scale grammar implementations*

While it makes good sense to work out our feature descriptions on the basis of samples of key data, it is also important to ask whether they ‘scale up’ when used in large-scale applications. There are indeed substantial projects, based on HPSG and LFG. Based on HPSG there is the CSLI LINGuistic Grammars Online (LINGO) project (<http://lingo.stanford.edu/>); this includes the English Resource Grammar (ERG) and the LKB (Lexical Knowledge Builder) grammar engineering system. For LFG there is the Parallel Grammar Project (ParGram), see <http://pargram.b.uib.no/>. This project includes a commitment to restrict the feature inventory, but the languages tackled so far cover a restricted typological space and so the features and values proposed are somewhat limited.

#### *2.4.7 Tagging a large corpus*

Similarly, feature sets have to be up to the task of tagging large corpora, including corpora of morphologically rich languages. The IPI PAN corpus of Polish has over a million words, and the requirements for tagging were challenging.<sup>11</sup> The difficulties and analytical choices made are described in Przepiórkowski (2004: 22-37). Such large projects, where a whole corpus has to be accounted for, provide stern tests for feature inventories.

### **2.5 Conclusion and prospects**

Features are central in mainstream linguistics, and they enjoy similar importance in linguistic frameworks which differ substantially in other respects. Features therefore need regular attention, so that we make principled rather than habitual choices in their use. Given their central position, it is not surprising that they bring with them many

issues which need debate and resolution. A bright prospect is the bringing together of research into the logic of features, in computational work and some theoretical work, with the work on the substantive semantics of features, which is mainly due to typologists. A second hopeful sign is the ‘bottom-up’ standardization initiated by the Leipzig Glossing Rules. It is evident that we should know what is intended by others’ glossing of examples. We should continue along this path, sharing definitions and conventions wherever possible, so that genuine theoretical differences are highlighted and evaluated.

## References

- Adger, David (2006). ‘Remarks on Minimalist feature theory and Move’. *Journal of Linguistics* 42: 663-673.
- Aristar Dry, Helen (2002). ‘E-MELD: Overview and Update’. International Workshop on Resources and Tools in Field Linguistics, Las Palmas 26 - 27 May 2002. [available at: <http://linguistlist.org/emeld/documents/index.cfm> ]
- Asudeh, Ash, and Toivonen, Ida (2006a). ‘Symptomatic imperfections’. *Journal of Linguistics* 42: 395-422.
- Asudeh, Ash, and Toivonen, Ida (2006b). ‘Response to David Adger’s “Remarks on Minimalist feature theory and Move”’. *Journal of Linguistics* 42: 675-686.
- Baerman, Matthew, Brown, Dunstan, and Corbett, Greville G. (2005). *The Syntax-Morphology Interface: A study of syncretism*. Cambridge: Cambridge University Press.
- Barlow, Michael (1992). *A Situated Theory of Agreement*. New York: Garland.  
[Published version of 1988 doctoral dissertation, Stanford.]

- Booij, Geert (1996). 'Inherent versus contextual inflection and the split morphology hypothesis', in Geert Booij and Jaap van Marle (eds) *Yearbook of Morphology 1995*. Dordrecht: Kluwer, 1-15.
- Carpenter, Bob (1992). *The logic of typed feature structures: With applications to unification grammars, logic programs and constraint resolution*. Cambridge: Cambridge University Press.
- Carpenter, Bob (2002). 'Constraint-based Processing', in Lynn Nadel (ed.) *Encyclopedia of Cognitive Science I*. London: Nature Publishing Group, 800-804.
- Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chumakina, Marina, Brown, Dunstan, Quilliam, Harley, and Corbett, Greville G. (2007). *Slovar' arčinskogo jazyka (arčinsko-anglo-russkij)* [A dictionary of Archi: Achi-Russian-English]. Makhachkala: Delovoj Mir. [More accessible is the WWW version at: <http://www.smg.surrey.ac.uk/> .]
- Chvany, Catherine V. (1986). 'Jakobson's fourth and fifth dimensions: on reconciling the cube model of case meanings with the two-dimensional matrices for case forms', in Richard D. Brecht and James Levine (eds) *Case in Slavic*. Columbus, OH.: Slavica, 107-129.
- Colmerauer, Alain (1970). 'Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur'. Internal publication 43, Département d'informatique de l'Université de Montréal, Septembre 1970.
- Comrie, Bernard, Haspelmath, Martin, and Bickel, Balthasar (2004). 'The Leipzig Glossing Rules'. Available at: <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>. [Revised 2008.]

- Copestake, Ann (1995). 'The representation of group denoting nouns in a lexical knowledge database', in Patrick Saint-Dizier and Evelyne Viegas (eds) *Computational Lexical Semantics*. Cambridge: Cambridge University Press, 207-230.
- Copestake, Ann (2002). *Implementing Typed Feature Structure Grammars* (CSLI lecture notes 110). Stanford: CSLI.
- Corbett, Greville G. (1987). 'The morphology/syntax interface: evidence from possessive adjectives in Slavonic'. *Language* 63: 299-345.
- Corbett, Greville G. (1991). *Gender*. Cambridge: Cambridge University Press.
- Corbett, Greville G. (2000). *Number*. Cambridge: Cambridge University Press.
- Corbett, Greville G. (2006). *Agreement*. Cambridge: Cambridge University Press.
- Corbett, Greville G. (2007). 'Canonical typology, suppletion and possible words'. *Language* 83: 8-42.
- Corbett, Greville G. (2008). 'Determining morphosyntactic feature values: the case of case', in Greville G. Corbett and Michael Noonan (eds) *Case and grammatical relations: papers in honor of Bernard Comrie* (Typological Studies in Language 81). Amsterdam: John Benjamins, 1-34.
- Corbett, Greville G., and Baerman, Matthew (2006). 'Prolegomena to a typology of morphological features'. *Morphology* 16: 231-246.
- Corbett, Greville G., and Hayward, Richard J. (1987). 'Gender and number in Bayso'. *Lingua* 73: 1-28.
- Dalrymple, Mary, and Kaplan, Ronald M. (2000). 'Feature indeterminacy and feature resolution'. *Language* 76: 759-98.
- Evans, Nicholas (1997). 'Head classes and agreement classes in the Mayali dialect chain', in Mark Harvey and Nicholas Reid (eds) *Nominal Classification in*

- Aboriginal Australia* (Studies in language companion series 37). Amsterdam: John Benjamins, 105-46.
- Farrar, Scott, and Langendoen, D. Terence (2003). 'A Linguistic Ontology for the Semantic Web'. *GLOT International* 7, no. 3.97-100
- Faßke, Helmut (1981). *Grammatik der obersorbischen Schriftsprache der Gegenwart: Morphologie*. Bautzen: Domowina Verlag.
- Franks, Steven (1995). *Parameters of Slavic Morphosyntax* (Oxford Studies in Comparative Syntax). New York: Oxford University Press.
- Fraser, Norman M., and Corbett, Greville G. (1997). 'Defaults in Arapesh'. *Lingua* 103: 25-57.
- Gazdar, Gerald, and Mellish, Chris (1989). *Natural Language Processing in Prolog: An Introduction to Computational Linguistics*. Wokingham: Addison-Wesley.
- Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey K., and Sag, Ivan A. (1985). *Generalized Phrase Structure Grammar*. Blackwell: Oxford
- Ginzburg, Jonathan, and Sag, Ivan A. (2000). *Interrogative Investigations: The form, meaning, and use of English interrogatives*. Stanford: CSLI.
- Greenberg, Joseph H. (1963). 'Some universals of grammar with particular reference to the order of meaningful elements', in Joseph H. Greenberg (ed) *Universals of Language*. Cambridge, MA: MIT Press, 73-113. [Paperback edition published 1966; page references to this edition.]
- Halle, Morris (1957). 'In defence of the number two', in Ernst Pulgram (ed.) *Studies Presented to Joshua Whatmough on His Sixtieth Birthday*. 's-Gravenhage: Mouton, 65-72.
- Haspelmath, Martin (2006). 'Against markedness (and what to replace it with)'. *Journal of Linguistics* 42: 25-70.

- van Helden, W. Andries (1993). *Case and gender: Concept formation between morphology and syntax (II volumes)* (Studies in Slavic and General Linguistics 20). Amsterdam: Rodopi.
- Jakobson, Roman O. (1958). 'Morfologičeskie nabljudenija nad slavjanskim sklonenijem (sostav russkix padežnyx form)', in *American Contributions to the Fourth International Congress of Slavists, Moscow, September 1958*. The Hague: Mouton, 127-156. [Reprinted in Roman Jakobson (1971) *Selected Writings II*. The Hague: Mouton, 154-183. Translated 1984 as: 'Morphological observations on Slavic declension (the structure of Russian case forms)', in Linda R. Waugh and Morris Halle (eds) *Roman Jakobson. Russian and Slavic grammar: Studies 1931-1981*. Berlin: Mouton de Gruyter, 105-133.]
- Kay, Martin (1979). 'Functional grammar'. *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistic Society*. Berkeley: BLS, 142-158.
- Kuhn, Jonas, and Sadler, Louisa (2007). 'Single conjunct agreement and the formal treatment of coordination in LFG'. Paper at LFG07, Stanford, 30 July 2007.
- Laidig, Wyn D., and Laidig, Carol J. (1990). 'Larike pronouns: duals and trials in a Central Moluccan language'. *Oceanic Linguistics* 29: 87-109.
- Lakoff, George (1972). Foreword in *Where the rules fail: a student's guide: an unauthorized appendix to M. K. Burt's 'From deep to surface structure'* [prepared by Susan Andres et al., rewritten and edited by Ann Borkin with the assistance of David Peterson]; with foreword by George Lakoff, pp. ii-v. Bloomington: Indiana University Linguistics Club.
- Leech, Geoffrey, and Wilson, Andrew (main authors) (1996). *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora (EAGLES*

Document EAG–TCWG–MAC/R). [available at:

[www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/annotate.ps.gz](http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/annotate.ps.gz)

- Lenček, Rado L. (1972). 'O zaznamovanosti in nevtralizaciji slovnične kategorije spola v slovenskem knjižnem jeziku'. *Slavistična revija* 20: 55-63.
- Levin, Magnus (2001). *Agreement with Collective Nouns in English* (Lund Studies in English 103). Stockholm: Almqvist & Wiksell.
- Matthews, P. H. (1972). *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*. Cambridge: Cambridge University Press.
- Meyer, Peter (1994). 'Grammatical categories and the methodology of linguistics: Review article on van Helden, W. Andries: 1993, *Case and gender: concept formation between morphology and syntax*'. *Russian Linguistics* 18: 341-377.
- Pensalfini, Robert (2003). *A Grammar of Jingulu: An Aboriginal language of the Northern Territory* (Pacific Linguistics 536). Canberra: Pacific Linguistics.
- Pollard, Carl, and Sag, Ivan A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
- Przepiórkowski, Adam (2004). *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.
- Pullum, Geoffrey K., and Zwicky, Arnold M. (1988). The syntax-phonology interface, in Frederick J. Newmeyer (ed.) *Linguistics: The Cambridge Survey: I: Linguistic Theory: Foundations*. Cambridge: Cambridge University Press, 255-80.
- Sag, Ivan A., Wasow, Thomas, and Bender, Emily (2003). *Syntactic Theory: A formal introduction*. Stanford: CSLI. [Second edition: first edition, Sag and Wasow 1999.]

- Saussure, Ferdinand de (1971). *Cours de linguistique générale* (publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger). Paris: Payot. [Third edition, first edition 1916.]
- Shieber, Stuart M. (1986). *An Introduction to Unification-Based Approaches to Grammar* (CSLI lecture notes 4). Stanford: CSLI, Stanford University,.
- Stanley, Richard (1967). 'Redundancy rules in phonology'. *Language* 43: 393-436.
- Stump, Gregory T. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.
- Stump, Gregory T. (2005). 'Word-formation and inflectional morphology', in Pavol Štekauer and Rochelle Lieber (eds) *Handbook of Word-formation* (Studies in Natural Language and Linguistic Theory 64). Dordrecht: Springer, 49-71.
- Svenonius, Peter (2007). 'Interpreting uninterpretable features'. *Linguistic Analysis* 33: 375-413.
- Warner, Anthony R. (1988). 'Feature percolation, unary features, and the coordination of English NPs'. *Natural Language and Linguistic Theory* 6: 39-54.
- Wechsler, Stephen, and Zlatić, Larisa (2003). *The Many Faces of Agreement*. Stanford: CSLI.
- Zaliznjak, Andrej A. (1973). 'O ponimanii termina 'padež' v lingvističeskix opisanijax', in Andrej A. Zaliznjak (ed.) *Problemy grammatičeskogo modelirovanija*. Moscow: Nauka, 53-87.
- Zwicky, Arnold M. (1986a). 'German adjective agreement in GPSG'. *Linguistics* 24: 957-990.
- Zwicky, Arnold M. (1986b). 'Imposed versus inherent feature specifications, and other multiple feature markings', in *The Indiana University Linguistics Club 20th Anniversary Volume*. Bloomington: Indiana University Linguistics Club, 85-106.

Zwicky, Arnold M. (1996). 'Syntax and phonology', in Keith Brown and Jim Miller (eds) *Concise Encyclopedia of Syntactic Theories*. Oxford: Elsevier Science, 300-305.

Table 2.1 Agreement with *committee* nouns (Levin 2001: 109)

	verb		relative pronoun		personal pronoun	
	<i>N</i>	% plural	<i>N</i>	% plural	<i>N</i>	% plural
US (spoken)	524	9	43	74	239	94
GB (spoken)	2086	32	277	58	607	72

Table 2.2 The Agreement Hierarchy: a sample of the evidence from gender

	attributive	predicate	relative pronoun	personal pronoun
Chichewa diminutive for human	gender 7	gender 7	gender 7	gender 7/ (GENDER 1)
Serbian/Croatian/Bosnian <i>d(j)evojče</i> 'girl'	n	n	n	n / F
Polish <i>tajdaki</i> 'wretches'	non_m.pers	non_m.pers / M.PERS	M.PERS	M.PERS
Konkani young females	f	N	no data	N
Russian <i>vrač</i> 'doctor' (female)	m / (F)	m / F	(m) / F	(m) / F
Serbian/Croatian/Bosnian <i>gazde</i> 'bosses'	f / (M)	(f) / M	((f)) / M	M

Notes: 1. lower case indicates syntactic agreement, and upper case SEMANTIC AGREEMENT

2. parentheses indicate a less frequent variant

Table 2.3 The initial stage in establishing features and values (abstract schema)

	Item1	Item2	Item3	...	...	...
Context1						
Context2						
Context3						
...						
...						
...						

Table 2.4 Establishing features and values: an example from Russian

	Item1 <i>žurnal</i> 'magazine'	Item2 <i>gazeta</i> 'newspaper'	Item3	...	...	...
Context1 <i>Na stole ležit ...</i> 'on table lies ...'	<i>žurnal</i>	<i>gazeta</i>				
Context2 <i>Ona думаet o ...</i> 'she thinks about ...'	<i>žurnale</i>	<i>gazete</i>				
Context3 <i>Ona čitaet ...</i> 'she reads ...'	<i>žurnal</i>	<i>gazetu</i>				
...						
...						
...						

---

<sup>1</sup> The support of the ESRC under grants RES-051-27-0122 and RES-062-23-0696 is gratefully acknowledged. I also wish to thank several colleagues for helpful comments: the participants in the Workshop on Features (1-2 September 2007), held in association with the LAGB meeting in London; Matthew Baerman, Dunstan Brown, Marina Chumakina, Patricia Cabredo Hofherr, Anna Kibort, Alexander Krasovitsky, Geoffrey Pullum and Claire Turner; and two anonymous referees.

<sup>2</sup> Another possibility is unary features. They are employed mainly in phonology; for discussion of an example of their use in syntax see Warner (1988).

<sup>3</sup> Jakobson's approach, particularly his analysis of case in Russian using three binary features, has been widely discussed and taken further; two examples are Chvany (1986) and Franks (1995: 41-55). There are three difficulties. First, there are further case values not covered by this analysis (see Zaliznjak 1973, Corbett 2008). Second the analysis is supported by an appeal to syncretism, but does not cover all the actual syncretisms (Baerman, Brown and Corbett 2005: 210). Third, as Gerald Gazdar points out (personal communication), there are 6720 possible ways to describe eight values using three binary features. In view of this, unless there are principled reasons for postulating particular binary features from the outset, it should not be taken as significant if there is an analysis using binary features which is partially successful.

<sup>4</sup> German is not unique here: for comparable examples from other languages see Baerman, Brown and Corbett (2005: 59-61).

<sup>5</sup> The marker *-rni* (where *rn* indicates a retroflex nasal) is indeed an exponent of the feminine, of the ergative, and of the focus, the latter being a recent development from the ergative (Rob Pensalfini, personal communication).

---

<sup>6</sup> A more entertaining account can be found on the website of the book:

<http://hpsg.stanford.edu/book/slides/>.

<sup>7</sup> Carpenter (2002) gives a fine overview of constraint-based approaches; he notes how the identities for which agreement is responsible have been one motivation in the development of computational work on language, starting with Colmerauer's early research (1970).

<sup>8</sup> As the English relative pronoun does not mark number, Levin checked his data and confirmed that singular verbs are normally found with *which*, and plural with *who*. He then counted relative pronouns as singular or plural on this basis, rather than establishing their number each time from the verb. Since relative *that* allows greater choice he included predicates of *that* within the predicate count. These decisions blur the picture somewhat, but there is information for recalculating and reinterpreting his results (2001: 32-33, 55-60).

<sup>9</sup> For a recent analysis of partial agreement in LFG, see Kuhn and Sadler (2007).

<sup>10</sup> Initial steps towards such an inventory can be found at:

<http://www.features.surrey.ac.uk/>.

<sup>11</sup> The acronym comes from the name of the host institution (in Polish), the Institute of Computer Science, Polish Academy of Sciences. See: <http://korpus.pl/>.