1 Why features?

Language has such a central place in our lives and in research that it is difficult to find an outside vantage point from which to achieve real understanding. As linguists, we attempt to do this by treating language as our object, while restricting the use of language as the tool. As a result of this approach, linguistics is in an exciting phase. Theories compete for overlapping segments of the research space. There is a sense of great achievement in some areas and equally of uncertainty about common goals. In this rapidly changing scene, one constant is the use of features. Fieldworkers, sociolinguists, computational linguists, syntacticians, working on spoken or on signed languages, all standardly use features. They are the key underpinning for linguistic description. We use features a good deal, but sometimes we take them for granted, assuming we all share the same conventions. In reality, the use of superficially similar notations sometimes hides differences in the underlying logic as well as in the substantive semantics of features.

It is therefore worth working through the motivation for using features, and the choices available to us. Naturally, different researchers make different choices; the important thing is that these should be reasoned choices, and that they should be made explicit. We shall give special attention to syntax and morphology, since it is in these components that the use of features requires the clearest argumentation. This is because these features do not have direct correspondences in meaning or sound, we have not such immediate evidence for them, and hence must justify their use with particular care. Having isolated the distinctions which we model using features, it is natural to typologize across them. As with all typology, we need to consider carefully whether we are comparing like with like, an issue which we discuss in §5.1.

1.1 Why do we use features?

Linguistic entities (such as words, phrases, and so on) have recognizable characteristics which can be 'factored out' and modelled with features. These features show consistency across entities, and to some extent across languages.[1]

[1] There have been claims that features like GENDER, NUMBER and TENSE are linked to a specific gene; for a sober assessment of this claim see Marcus & Fisher (2003).

They allow us to say, for example, that within a given language the same distinctions of **NUMBER** occur across different constructions (agreement within the noun phrase as opposed to within the clause) and yet are realized differently across lexemes (thus *this : these :: runs : run*). They have become such a key part of the intellectual infrastructure that we sometimes hardly notice them – rather like electricity and water. A power failure can be helpful in reminding us what life would be like without electricity, and time at a field site without running water is also a salutary experience. So let us briefly try to do linguistics without features to see how central they are. We shall try to write a grammar of a small fragment of a language, not using features. We shall find quite quickly that without features we would be missing the point.

1.1.1 Generalizations in syntax

We shall see how features allow us to capture generalizations in syntax. We begin by writing a grammar for a fragment of Russian, without using features. For now I shall leave out interlinear glosses, since to include them we would need to know precisely what we are trying to establish.

(1) Russian
devuška pišet
'The girl is writing.'

Russian has no article: *devuška* simply means 'girl'. And *pišet* can convey both 'writes' and 'is writing'. We expand the fragment slightly:

(2) devuški pišut
'The girls are writing.'

(3) mal´čik pišet
'The boy is writing.'

(4) mal´čiki pišut
'The boys are writing.'

The remaining combinations are ungrammatical, for example:

(5) *devuška pišut
'The girl are writing.'

We can write a grammar to account for this tiny fragment of Russian, using PATR-II notation (following Gazdar & Mellish 1989: 218–20 on French).[2] PATR-II is a computer language, designed for computational linguistic purposes. Our grammar has two rules of syntax, and a small lexicon ('`pos`' indicates part of speech):

[2] A similar argument for features is given in Sag, Wasow & Bender (2003: 38–40), and in Bird, Klein & Loper (2009: 327–57), where an implementation in Python is given.

(6) Basic grammar of Russian

```
SYNTAX:
Rule
 S → A B.

Rule
 S → C D.

LEXICON:
Word devuška:
 <pos> = A.

Word devuški:
 <pos> = C.

Word mal´čik:
 <pos> = A.

Word mal´čiki:
 <pos> = C.

Word pišet:
 <pos> = B.

Word pišut:
 <pos> = D.
```

This toy grammar successfully generates all and only the sentences in our fragment. S is rewritten as AB or CD according to the rules in the syntax, and then appropriate items are chosen as A, B, C or D according to their part of speech (pos) labels in the lexicon. We could easily add more words to our lexicon, and the coverage would increase accordingly. However, this toy grammar is not very exciting or insightful; it is evident that we are missing something. In fact we are missing the same thing twice. First, in terms of the syntax, our two rules both state that Russian has intransitive sentences. If we extend our grammar to include transitive sentences, we shall have to add two rules (to allow singular and plural direct objects), and continue adding two more for each such extension. Second, in terms of the lexicon, each time we add a new word, we are likely to have to make two new entries. All this suggests that our analysis is uneconomical, lacking insight and, well, rather boring.

To remedy this, we want to 'factor out' the feature **NUMBER**. This would allow us to say that in our syntax, A and C are essentially the same, and equally that B and D are essentially the same. We could call them AC and BD; to give our grammar a more familiar look, we will instead give them syntactic category labels NP and VP. We can then use more recognizable syntactic rules:

(7) Grammar of Russian: version 2 (syntax only)

SYNTAX:
Rule
S → NP_{sg} VP_{sg}.

Rule
S → NP_{pl} VP_{pl}.

Our grammar now reflects the fact that we have been dealing with only two types of syntactic object, both of which differ according to **NUMBER**. We now have complex symbols (since for instance NP_{sg} consists of NP plus $_{sg}$) but this does not lead to an increase in expressive power. The reason is that we can refer to the specifications (such as 'pl') to make generalizations, but equally we could interpret the symbol NP_{pl} as a single ornate symbol (treating it as the equivalent of C above), see Halle (1969) and Gazdar, Klein, Pullum & Sag (1985: 20–1). Thus introducing features abbreviates the grammar but does not change the expressive power of a grammar (Coleman 1998: 105–8). This means that features have the advantage of allowing us to capture generalizations without making our theory less restricted. Of course, we want our theory to be as simple (restricted) as possible while covering as much of the observed data as possible. Nevertheless, we still have two rules in (7) which are essentially saying the same thing.

We should go further, therefore, and separate the featural information from the structural rule (see again Gazdar & Mellish 1989: 219):

(8) Grammar of Russian: version 3 (syntax only)

SYNTAX:
Rule
S → NP VP
<NP number> = <VP number>.

Here we have a single syntactic rule for intransitive sentences; we have a constraint on it, namely that the number of the NP and that of the VP must match (where **NUMBER** has the possible values SINGULAR and PLURAL). That is, we have a structural rule and an agreement rule, which is stated as a constraint. (We also need to add rewrite rules to rewrite NP as A or C and VP as B or D.) For a more technical account see Sag, Wasow & Bender (2003: 69–72, 107–118).

1.1.2 Generalizations in morphology

Just as the original syntactic rules in our mini-grammar missed the point, by ignoring the regularity of **NUMBER**, so did the original lexical entries. Recall that we had these four entries (we will concentrate on nouns, though an analogous argument can be made with verbs):

(9) Basic Russian morphology

devuška 'girl'	category A
mal´čik 'boy'	category A
devuški 'girls'	category C
mal´čiki 'boys'	category C

While it is possible to treat these items as belonging to separate categories, there is no need to do so, now that we have a more natural syntax. It would make sense to factor out **NUMBER** in the morphology too, and to treat the four items as belonging to the same part of speech (lexical category), namely N for noun. That is, we recognize that their part of speech is the primary classification, and that the **NUMBER** feature gives a secondary classification. We now treat each item as having its lexical meaning, stated simply here as 'boy' or 'girl', and its grammatical meaning SINGULAR or PLURAL. We classify the items as follows:

(10) Russian morphology: version 2

SINGULAR	PLURAL	Gloss
devuška	devuški	'girl(s)'
mal´čik	mal´čiki	'boy(s)'

A natural conclusion is to say that we are dealing with two lexemes (which we may label by the citation forms, the SINGULAR *devuška* and *mal´čik*) each of which has two inflectional forms, SINGULAR and PLURAL (we will return to the 'pos' features shortly):

(11) Lexicon for Russian morphology: version 2

```
LEXICON:
Lexeme devuška:
 <pos> = N
 <sg> = devuška
 <pl> = devuški.

Lexeme mal´čik:
 <pos> = N
 <sg> = mal´čik
 <pl> = mal´čiki.
```

Though this is a tiny example, it contains the key insight which we wish to capture through the use of features. Let us glance back to (10), and ask which items are similar. In an obvious way, *devuški* 'girls' is very like *devuška* 'girl'; they share their lexical meaning – they are forms of the same lexeme. From a grammatical point of view, however, *devuški* 'girls' is more like *mal´čiki* 'boys' since they can both fit into similar slots in a sentence: they share a grammatical meaning (PLURAL). Features allow us to capture these cross-cutting classifications, and this is a powerful argument in their favour. Thus they allow us to model similarities

which are not full identities. They allow us to treat all the PLURAL nouns as a class, just as in phonology features make it possible to pick out the 'natural classes' of segments.[3]

This solution in (11) is all the more convincing in languages which have additional number values, such as DUAL and PAUCAL, since then it would save us having to have three or four lexical entries for each noun. The step we have taken may seem obvious and innocent, but it has implications. Note that though we say that each of our two nouns has a SINGULAR and a PLURAL, the distinction is marked differently. The first marks the singular with stem + *a*, and the second by the bare stem (as in English).[4] Both show similar plurals, but there are other Russian nouns which differ. This is another advantage of the use of features. We want to say that these two nouns mark the same distinction, even though they realize it differently.

In (11) I slipped in a further development in our use of features. The **PART OF SPEECH** (pos) is also treated as a feature. Parts of speech are rather different from the features we have discussed so far. In the simplest instance, an item has a single part of speech specification, say VERB, but various possible values for **TENSE**, **PERSON, NUMBER** and so on. And yet, each feature is used to divide up a set of linguistic elements. Features like **CASE** and **NUMBER** cross-classify, as we shall see in §1.2. Similarly, if we treat **PART OF SPEECH** as a feature, with values like NOUN, VERB, ADJECTIVE, we can classify words, and this classification may cross-cut others, particularly **NUMBER** as we have just observed (see further in §3.7). Since the morphosyntactic features show great diversity and have generally been studied less well than parts of speech, we shall give them particular attention.[5]

In the examples in (10) and (11), lexical and grammatical meaning combine in a compositional way. The whole is the predictable sum of the parts. That is, *devuški* 'girls' = *devuška* + PLURAL. And the following relation holds: *devuški* is to *devuška* as *mal´čiki* is to *mal´čik*. We expect the difference in meaning between *devuški* 'girls' and *mal´čiki* 'boys' to be entirely due to their lexical meaning, with plurality remaining constant. This is true in the canonical instances, but it is not always so (see §8.3). To take just one example: Russian *nožnicy*, like its English translation 'scissors', is a plurale tantum noun. Here PLURAL is not equivalent to PLURAL with *devuški* 'girls', since *nožnicy* can perfectly well be used of one pair of scissors (there is more on such nouns in §8.1). Why then label *nožnicy* as PLURAL? We do this, because it takes the same form of the verb as

[3] Natural classes were discussed in phonology, where the naturalness is a matter both of segments being subject to similar rules (as when voiced consonants are devoiced) and being phonetically natural too (Postal 1968: 73–5, Gussenhoven & Jacobs 2005: 58); for discussion see Spencer (1996: 130–8). But outside phonology too, features equally pick out natural classes in their relevant domain; an early and interesting illustration of this is Bierwisch (1967).

[4] The feature value is a description of the whole form, and cannot be associated just with the affix.

[5] It is also worth noting a problem of linguists' usage. While it is readily accepted that parts of speech, inflectional classes, and so on, can be modelled with features, many linguists take 'feature' to mean 'morphosyntactic feature'. There is the assumption that these are the 'real' features in some sense.

do plural nouns, which was our original motivation for introducing the NUMBER feature. As we shall discuss in §5.3, NUMBER is a morphosyntactic feature (it is relevant for both syntax and morphology) and its relation to semantics is not always straightforward.

Once we have factored out the NUMBER feature, this opens up a range of typological questions. We have touched on the question as to whether a feature value like PLURAL always means the same thing, and that has cross-linguistic implications (see §5.1.1 for the general issue, the correspondence problem, and §5.3 specifically for NUMBER). We may also ask whether the same inventory of lexemes mark NUMBER. In fact there is massive cross-linguistic variation here; for instance, there are languages where almost every noun marks NUMBER, and others where NUMBER is restricted to a very few (Corbett 2000: 54–75).

We have used the feature NUMBER, with the values SINGULAR and PLURAL, on both nouns and verbs. For English and Russian this makes excellent sense, since the system is similar for both. In some languages there is no such straightforward match (see §5.1.2 and §8.1). But whether there is a match or not, we still have to ask whether the feature is the same across parts of speech. We return to this problem in §3.6.

1.2 Orthogonal features (in syntax and in morphology)

The picture of Russian given so far has been simplified, and when I make it a little more realistic we see again a powerful reason for the use of features. The forms of the noun we have considered are actually those of the NOMINATIVE, which is just one case value out of several. Here are fuller paradigms:

(12) Paradigms of Russian *devuška* 'girl' and *mal´čik* 'boy'

	SINGULAR	PLURAL	SINGULAR	PLURAL
NOMINATIVE	devuška	devuški	mal´čik	mal´čiki
ACCUSATIVE	devušku	devušek	mal´čika	mal´čikov
GENITIVE	devuški	devušek	mal´čika	mal´čikov
DATIVE	devuške	devuškam	mal´čiku	mal´čikam
INSTRUMENTAL	devuškoj	devuškami	mal´čikom	mal´čikami
LOCATIVE	devuške	devuškax	mal´čike	mal´čikax

These paradigms are laid out as we might find them in a pedagogical work. But theoretical linguists use the same system. It is important to be clear about what is claimed. First the layout implies that we are giving forms of the same lexeme, where each cell consists of a combination of lexical and grammatical meaning. The layout reflects the claim that NUMBER (a binary feature in Russian), cross-cuts with CASE, which has several values (I give the basic six in (12), and discuss the full range in Chapter 7). Had we tried to maintain the original syntactic analysis

not using features, the effect of these case distinctions on the syntactic rules would have been dramatic. By using two orthogonal (cross-cutting) features, we allow the syntax to refer to specific case values (e.g. particular verbs govern the ACCUSATIVE, GENITIVE, DATIVE or INSTRUMENTAL) and to refer to number values (as in agreement, as we saw earlier).[6] And the features can be found elsewhere in the description: I present nouns here, but adjectives, pronouns and numerals also have case distinctions.

We shall return to paradigms like those in (12), but here it is worth unravelling more of the assumptions involved. Forms which are phonologically different, for example the inflected forms in the first cell for each noun, one in *-a* and one with the bare stem, are given the same description, namely NOMINATIVE SINGULAR. Just as different forms are given the same description, so the same forms can have different descriptions; for instance, in the second paradigm the form *mal´čika* is both ACCUSATIVE SINGULAR and GENITIVE SINGULAR of *mal´čik* 'boy'. This is an instance of **syncretism**, the use of a single form for more than one function, which is something we return to in §2.3. The assumption which prompts both these non-obvious mappings between the forms and their specifications is the principle that syntax is 'morphology-free'. We aim for simple rules of syntax, referring to featural specifications such as ACCUSATIVE, not rules which have access to the way in which such specifications are realized for particular nouns. We consider this more fully in §3.4.8. More generally, the two nouns do not have the same number of distinct forms, yet they are fitted into the same shape of paradigm; clearly we shall need to justify the claim that they each mark six case values. We discuss this in detail in Chapters 4 and 6.

Before leaving (12), note that the ordering of the case values is a matter of tradition and convenience; no specific claim is being made about possible relations between case values by their relative position in the table.

1.3 Practical issues

It is worth considering practical issues at this point, since we can here adopt conventions which will inform the rest of the book. When presenting texts, whether larger texts or small examples, linguists normally provide featural information to help the reader. This may be as a minor aid to someone who is reading the text for a quite different purpose, it may permit the reader to focus on some other linguistic point about the examples, or the featural information may be the main point of the example (as it typically will be for us). For this we shall adopt the Leipzig Glossing Rules (available at www.eva.mpg.de/lingua/resources/glossing-rules.php). These conventions were put forward by Comrie,

[6] While in §1.1.2 we saw **NUMBER** cross-cutting with **PART OF SPEECH**, having **NUMBER** cross-cutting with **CASE** within individual lexemes as in (12) makes the point if anything more strongly.

Haspelmath & Bickel (2004), following Lehmann (1983). They were revised in 2008. The rules include a 'lexicon', that is a set of abbreviations for feature values.[7] We shall use these, and add to the list when necessary (the full list of abbreviations used is on pages xvi–xviii). We give the basic principles of the Leipzig Glossing Rules here, and we leave the reader to consult the original for further detail, especially for additional optional rules.

The basic layout of a glossed example is as in (13):

(13)
Russian (the source may be given here)
devuški pišut [object language]
girls write [interlinear gloss]
'The girls are writing.' [translation]

While some give the language before each example, this information is often 'carried forward': the following example is assumed to be from the same language, as with (14) below, unless it is labelled otherwise. Note that the gloss is left-aligned vertically, word by word, with the object language example. It matches the object language one-for-one (we shall come to instances where this is not straightforward below). The translation, on the other hand, gives an indication of the meaning. Thus though Russian has no articles, the definite article is included in the translation of (13) to give the best indication of the meaning of the source language example.

We can give the reader more information, however. The object language words can be segmented, and the interlinear gloss can include featural information:[8]

(14)
devušk-i piš-ut [object language]
girl-PL write-PL [interlinear gloss]
'The girls are writing.' [translation]

Both words can be segmented into stem and affix, and we mark this segmentation with hyphens. The lexical meaning is given as before, and the value of the feature in small capitals. The feature value is often abbreviated, as here; either abbreviations are taken from the standard list, or are specified. The glossing conventions require that there should be the same number of hyphens in the example as in the gloss. When clitics are involved, rather than affixes, these are marked off with '='.

There is further information that we could give. As we saw in (12), *devuški* is the NOMINATIVE PLURAL. And *pišut* 'write' is not available for all persons in the

[7] The list in the Leipzig Glossing Rules is simply a list: the values are not typed (see §2.4). That is, the list includes 'PL' as an abbreviation for PLURAL, but it does not indicate that it is available as a feature value of **NUMBER**, and not of **CASE** or **GENDER** (see further in the initial part of Chapter 8).

[8] It is also possible to give the featural information without having to commit to any segmentation. Thus *devuški* can be glossed as girl.PL or indeed as girl.PL.NOM without segmentation. Some researchers provide an additional line, giving the example 'straight' and then with segmentation. One example of this is given in the book: see example (18) in §5.4.

PLURAL, only the THIRD person PLURAL. Thus a more detailed version would be the following:

(15) devušk-i piš-ut [object language]
girl-PL.NOM write-3PL [interlinear gloss]
'The girls are writing.' [translation]

Consider the glossing of the *-i* ending of *devušk-i* 'girls'. This is an instance of a single object-language element being glossed by more than one metalanguage element. We need to reflect the fact that the *-i* marks both PLURAL and NOMINATIVE. The stop (period) is used in PL.NOM to indicate this, and to preserve the one-to-one match between the segmented elements in the source language and the gloss.[9] The glossing of the verb also deserves attention. By convention, there is no separating marker between **PERSON** and **NUMBER** (hence '3PL'), when they co-occur in this order. Here tradition outweighs consistency (which would have required '3.PL').

It is notable just how many significant issues arise in the apparently innocent business of providing glosses for a couple of rather anodyne examples. We have glossed *devuški* 'girls' as NOMINATIVE PLURAL. However, the paradigm given in (12) shows that it could in principle be GENITIVE SINGULAR. Though the decision is rarely made explicit, linguists normally provide what is actually a morphosyntactic gloss. That is to say, we consider the syntactic context and use it to resolve potential morphological ambiguities. Given an example like *We bid ninety pounds for a donkey yesterday*, most linguists would gloss *bid* as PAST, even though out of context it could be the PRESENT or even the IMPERATIVE. Following the same principle, in (15) *devuški* 'girls' was glossed as NOMINATIVE PLURAL since this provides a normal, grammatical sentence of Russian. And this, of course, rests on a set of analyses: Russian typically has NOMINATIVE subjects and treating *devuški* 'girls' as NOMINATIVE PLURAL gives an acceptable sentence. If we treated it as GENITIVE SINGULAR, there would be two problems: while GENITIVE subjects are possible, this is not so in this sentence type; and verbs do not agree with GENITIVE subjects. The alternative analysis (GENITIVE SINGULAR) simply will not work. For most purposes, this background may not be necessary, and so it is reasonable to take the short cut and give a morphosyntactic gloss. For some purposes, however, such as in a discussion of syncretism, theoretical ambiguity might be discussed explicitly in the text, or the gloss could include all the possibilities: GEN.SG / NOM.PL. Note that we are using the slash for alternative glosses; it is available for this use because the other function it may fulfil, namely

[9] Since both PL and NOM are equally required, it would be reasonable to treat the ordering of these elements as unimportant, and many researchers do that. Alternatively one can import more of one's theory into the glossing. I would argue that **NUMBER** is more relevant to the noun than is **CASE**, and that the ordering should respect this (hence 'girl-PL.NOM'). The greater relevance of **NUMBER** is shown by the fact that some Russian nouns have different stems for SINGULAR and PLURAL, irrespective of **CASE**, but no noun has different stems for **CASE**, irrespective of **NUMBER**. Where possible, I shall use a principled order.