1 Word: a typological framework

R. M. W. Dixon and Alexandra Y. Aikhenvald

In this book we ask how 'word' should be defined. What are the criteria for 'word'? Is 'word', as the term is generally understood, an appropriate unit to recognise for every type of language?

This introductory chapter first looks at what scholars have said about 'word', and then discusses the categories and distinctions which need to be examined. Chapter 2 suggests a number of typological parameters for the study of clitics. Following chapters then provide detailed examination of the notion of 'word' in a selection of spoken languages from Africa, North and South America, Australia, the Caucasus and Greece, together with a discussion of words in sign languages. The final chapter, by P. H. Matthews, asks what has been learnt from these general and particular studies.

This introduction begins by surveying the criteria that have been put forward for 'word', and suggests that one should sensibly keep apart phonological criteria, which define 'phonological word', and grammatical criteria, which define 'grammatical word'. In some languages the two types of word coincide and one can then felicitously talk of a single unit 'word', which has a place both in the hierarchy of phonological units and in the hierarchy of grammatical units. In other languages phonological word and grammatical word generally coincide, but do not always do so. We may have a grammatical word consisting of a whole number of phonological words, or a phonological word consisting of a whole number of grammatical words. Or there can be a more complex correspondence between the two types of word with, say, a grammatical word consisting of all of one and part of another phonological word.

§1 summarises the tradition, §2 discusses linguists who would do without the word and §3 surveys opinions concerning 'what is a word'. In §4 a number of confusions are discussed and then in §5 some suggested criteria are examined. The heart of the chapter is in §§6–8 – proposed definitions for phonological word and for grammatical word (and the status of clitics) – followed by (in §9) examination of the relationship between the two types of word. In §10 we ask whether all kinds of languages have words; in §11 there is brief discussion of the varying social status of 'word' in different languages, and then §12 provides a summary of the results of the introductory chapter. Finally, the appendix gives a

brief statement of the criteria for phonological word and for grammatical word (and their relationship) in a sample language, Fijian.

1 The tradition

Many writers have assumed that 'word' is a – or the – basic unit of language. Bolinger (1963: 113) comments: 'Why is it that the element of language which the naive speaker feels that [they] know best is the one about which linguists say the least? To the untutored person, speaking is putting words together, writing is a matter of correct word-spelling and word-spacing, translating is getting words to match words, meaning is a matter of word definitions, and linguistic change is merely the addition or loss or corruption of words.' Bolinger himself takes 'word' as a prime, commenting that it is 'the source, not the result, of phonemic contrasts'.

And, as Lyons (1968: 194) comments: 'The word is the unit *par excellence* of traditional grammatical theory. It is the basis of the distinction which is frequently drawn between morphology and syntax and it is the principal unit of lexicography (or "dictionary-making")'. Indeed, for the Greeks and Romans the word was the basic unit for the statement of morphological patterns; they used a 'word and paradigm' approach, setting out the various grammatical forms of a given lexeme in corresponding rows and columns, with no attempt to segment into morphemes (Robins 1967: 25). (In fact Greek and Latin are fusional languages where it is not an easy matter to segment words into morphemes, without bringing in the impedimenta of underlying forms, morphophonological rules, and the like.)

Much that has been written about the word is decidedly eurocentric. It has sometimes been said that 'primitive languages' do not have words, an opinion which Lyons (1968: 199) explicitly rejects, partly on the basis of Sapir's report that uneducated speakers of American Indian languages can dictate 'word by word'.

However, it appears that only some languages actually have a lexeme with the meaning 'word'.[1] Even in some familiar languages where this does occur it may be a recent development. For instance, in Old English the primary meaning of *word* was (a) for referring to speech, as contrasted with act or thought. There was a second sense, which may then just have been emerging: (b) what occurs between spaces in written language. In the development to Modern English (b) has become the major sense – the one used in this book – with sense (a) still surviving mainly in fixed phrases, e.g. *the spoken word*,

[1] Dixon (1977a: 88) states 'every (or almost every) language has a word for "word"'; this is erroneous. Wierzbicka (1996, 1998) has 'word' as a universal semantic primitive, which is said to be realised in every language; this is equally erroneous.

the written word, *the Word of God*, *a word of warning/advice/caution*, *Can I have a quick word with you?* Similar remarks apply to corresponding terms in some other European languages, e.g. *mot* in French and *slovo* in Russian. German also has a noun, *Wort*, with these two senses, but there are here two plural forms – *Worte* for 'speech' and *Wörter* for 'what is written between spaces'.[2]

Over on the other side of the world, *vosa* in Fijian is a verb meaning 'speak, talk' and also a noun, with several related senses: 'language', 'talk, speech' and 'word'. It is likely that we have here a similar line of semantic development to *word* in English.

The vast majority of languages spoken by small tribal groups (with from a few hundred to a few thousand speakers) have a lexeme meaning '(proper) name' but none have the meaning 'word'. This applies to many languages from Australia (including Arrernte, chapter 4 in this volume), Amazonia (including Jarawara, chapter 5) and New Guinea with which we, or our colleagues, are familiar.

2 Doing without 'word'

The idea of 'word' as a unit of language was developed for the familiar languages of Europe which by-and-large have a synthetic structure. Indeed – as will be shown below – some of the criteria for 'word' are only fully applicable for languages of this type.

What about languages from extreme ends of the typological continuum – those of an analytic or of a polysynthetic profile? Reviewing the first edition of Nida's (1944) *Morphology*, Hockett (1944: 255) notes that Nida 'devotes a chapter to the criteria by which words may be recognised. None of these criteria, nor any combination of them, gives any fruitful results with Chinese . . . the real implication is that THERE ARE NO WORDS IN CHINESE. The whole tradition of "words" as worked out with western languages is useless in Chinese.' (However, a quite different opinion is expressed by the leading Chinese linguist, Chao, discussed in §10 and §11 below.)

Some of the polysynthetic languages of North America lack any unit that looks like the sort of word we are used to from European languages. Gray (1939: 146) presents his own definition of word as 'a complex of sounds which in itself possesses a meaning fixed and accepted by convention'. (Note that this would, in fact, also be satisfied by a prefix such as *un-* or a phrase such as

[2] It is likely that all languages with an established (non-ideographic) orthographic tradition do have a word for 'word'. Other languages tend to create such a term once they are exposed to writing. The interesting question is how many languages with no written tradition have a lexeme which corresponds to *word* in English, *mot* in French, etc.

The King of England.) Then, in a footnote, he reports a suggestion from H. J. Uldall that in polysynthetic languages (such as Maidu from north-east California) 'the word as such is not a relevant part of analysis'. This idea is repeated by Milewski (1951), who argues that in all languages of the world there are three kinds of morphological unit, morphemes, syntactic groups and clauses. However, Krámský (1969: 74–5) suggests that Milewski's 'syntactic group' in polysynthetic North American languages does satisfy criteria for 'word'. It is clear that the unit 'word' can be recognised in polysynthetic languages, it is just that it is much longer and more complex than the words linguists were used to at that time.

It is, however, important to distinguish between the structure of the predicate and the structure of the verb. This is discussed in §§5–6 of chapter 5, where it is shown that a language may have fairly complex predicate structure but relatively simple verb structure (as in English and Fijian), or simple predicate structure combined with somewhat complex verb structure (as in Dyirbal), or both complex predicate structure and complex verb structure (as in Jarawara).

There are those who consider 'word', as a general notion, not to be a basic category of language. The anthropologist Malinowski (1966: 11) insists that one should analyse utterances, not any smaller units of language taken out of their context of use. He can then say 'isolated words are in fact only linguistic figments, the products of an advanced linguistic analysis'.[3]

Other linguists accord a grudging role to 'word'. For Börgström (1954: 276) 'words are utterance-segments consisting of one or more morphemes. Assuming that there is a procedure for the demarcation of morphemes, it is possible, I believe, to formulate a set of distribution rules as a procedure for the demarcation of words.' Writing in the same year, Garvin (1954: 345) is less sure about this: 'in the present state of our techniques one may assume that we know how to isolate morphemes properly – that is, unequivocally and without unaccountable residue. It is not so certain that we know how to isolate words, and hence how to separate morphology from syntax.'

The extreme position is taken by Harris (1946: 161) who presents a procedure (illustrated for English and Hidatsa) for analysing utterances into morphemes: 'the method described in this paper will require no elements other than morphemes and sequences of morphemes, and no operation other than substitution, repeated again and again'. The unit 'word' does not feature in Harris' analysis.[4]

[3] We also find (perhaps as a further reflection of Malinowski's position) Potter's (1967: 78) statement: 'unlike a phoneme or a syllable, a word is not a linguistic unit at all. It is no more than a conventional or arbitrary segment of utterances.'

[4] We have noted one instance of *word* in this paper, but this is used in an informal rather than in an analytic sense. On page 166 Harris is discussing the English sentence *I know John was in* and talks of 'pronouncing its intonation twice, once over the first two words and again over the last three'.

3 What is a word?

Matthews commences the section 'What are words?' in the second edition of his seminal textbook *Morphology* (1991: 208) with: 'there have been many definitions of the word, and if any had been successful I would have given it long ago, instead of dodging the issue until now'.

Matthews mentions that the ancient grammarians simply had word as the smallest unit of syntax. But, he comments, to follow that line 'will only turn our larger problem back to front. If words are to be defined by reference to syntax, what in turn is syntax, and why are syntactic relations not contracted by parts of words as well as whole words?'

Some of the definitions suggested for word are horrifying in their complexity and clearly infringe the principle that a definition should not be more difficult to understand than the word it purports to define.[5] There are useful surveys of definitions of 'word' in Rosetti (1947), Weinreich (1954), Ullmann (1957) and Krámský (1969).

Some definitions are simple and appealing. These include Sapir's (1921: 34) 'one of the smallest, completely satisfying bits of isolated "meaning" into which the sentence resolves itself' and Žirmunskij's (1966: 66): 'the word is the most concise unit of language, which is independent in meaning and form'. But each of these is essentially vague; they do not provide definite criteria for deciding 'what is a word' in a given language.

Sweet (1875/6: 474) suggests: 'we may, therefore, define a word as an ultimate or indecomposable sentence'. That is, anything which is a word can make up a complete sentence. Sweet offers as examples of this (from English) *Come!* and *Up?* (meaning 'Shall we go up?'). However, he is then concerned over what to do with forms like English *the* and *a*, which he terms 'half-words'.

Bloomfield (1933: 178) pursues a similar line in his definition: 'a word, then, is a free form which does not consist entirely of (two or more) lesser free forms; in brief, a word is a *minimum free form*' (his italics). This is probably the most oft-quoted definition of 'word' but it is, in fact, scarcely workable. There is further discussion in §5 below.

[5] We can quote two rather extreme examples. Firstly, Longacre's (1964: 101) definition, which was conceived within the formal framework of tagmemics: 'a class of syntagmemes of a comparatively low hierarchical order, ranking below such syntagmemes as the phrase and the clause and above such syntagmemes as the stem (as well as above roots which have no external structure and are therefore not syntagmemes). It may be of greatly varied structure . . . Words tend to be rigidly ordered linear sequences containing tagmemes which (aside from those manifested by stems) are manifested by closed classes of morphemes unexpandable into morpheme sequences and giving only stereotyped bits of information.'

Krámský devotes a whole monograph to discussing 'word'. He surveys past definitions and then comes up with his own (1969: 67): 'the word is the smallest independent unit of language referring to a certain extra-linguistic reality or to a relation of such realities and characterised by certain formal features (acoustic, morphemic) either actually (as an independent component of the context) or potentially (as a unit of the lexical plan)'.

4 Confusions

The word 'word' is used in many ways in everyday speech, and in much linguistic discourse. It is important to make certain fundamental distinctions:

(1) between a lexeme and its varying forms;
(2) between an orthographic word (something written between two spaces) and other types of word;
(3) between a unit primarily defined on grammatical criteria and one primarily defined on phonological criteria.

These are discussed, in turn, in §§4.1–3.

The (grammatical) word forms the interface between morphology and syntax. Morphology deals with the composition of words while syntax deals with the combination of words. One could imagine slightly different words being required as ideal units for these two purposes. That is, there could be a 'morphological word' and a 'syntactic word' which would perhaps generally coincide but might not always do so. We are not aware of this sort of distinction having been fully justified for any language;[6] but it is certainly a possibility. (In chapter 7, Rankin et al. put forward the idea that the term 'syntactic word' could perhaps be used – in Siouan languages – for a type of word incorporating a relative clause, the whole constituting one phonological word.)

4.1 Word and lexeme

Consider the following examples, from English and Latin, of the root or underlying form of a lexeme and its inflected forms, as used in a sentence.

	root or underlying form	inflected forms	
(a)	*look*	*look*	present, non-3sg subject
		looks	present, 3sg subject
		looked	past
		looking	participle
(b)	*lup-* 'wolf'	*lupus*	nominative sg
		lupō	dative/ablative sg
		lupī	genitive sg, nominative pl
		etc.	

[6] The possibility of this is mentioned by Di Sciullo and Williams (1987) without, however, the formulation of any explicit cross-linguistic or language-specific criteria. This question is also aired in Gak (1990). Dai (1998) establishes separate units 'syntactic word', 'phonological word', and 'morphological word' in Chinese. He suggests that a compound is one syntactic word and also one morphological word but that it may have different syntactic and morphological structures.

A number of other types of 'word' have been suggested. For example, Packard (2000: 7–14) lists: orthographic word, sociological word, lexical word, semantic word, phonological word, morphological word, syntactic word, and psycholinguistic word.

The term 'word' is sometimes used in reference to the root or underlying form, and sometimes in reference to the inflected forms. That is we hear, on the one hand things like *'look*, *looks*, *looked* and *looking* are forms of the same word', and on the other hand things like 'the lexeme *look* is realised as word-forms *look*, *looks*, *looked* and *looking*'.

Bally (1950: 287–9) is so concerned about this ambiguity of usage that he recommends abandoning the label 'mot' in French (and 'word' in English) and instead employing 'sémantème' for the root or underlying form and 'molécule syntaxique' for inflected forms. Lyons (1968: 197) prefers a different course. While recognising that in classical grammar 'word' was used to mean 'sémantème' he notes that modern usage tends to employ 'word' as a label for 'molécule syntaxique' and suggests standardising on this.

We have followed Lyons' suggestion, of using 'lexeme' as the label for 'root or underlying form' and '(grammatical) word' for 'inflected form of a lexeme'. Note that Lyons uses italics for words and capitals for lexemes – thus, the word *looked* is the past tense form of the lexeme LOOK.

Lyons' convention is useful from another viewpoint, for dealing with lexemes that involve two words. These include phrasal verbs in English such as MAKE UP, as in *I made the story up* and *I made it up*. Note that the words of this lexeme are mapped onto two non-contiguous syntactic slots – an inflected form of *make* goes into the verb slot while *up* follows the object NP.[7] That is, the lexeme MAKE UP consists of two words, each of which has its own syntactic behaviour. If we had decided on 'word' as the label for lexeme, there would then be need for a separate notion of 'syntactic word'. We would have had to say that the (lexical) word *make up* consists of two syntactic words, *make* and *up*. This is avoided by describing MAKE UP as a lexeme that consists of two (grammatical) words, an inflected form of *make* and the preposition *up*. (Similar remarks apply to phenomena such as separable preverbs in German and Hungarian.)

4.2 Orthographic word

In many language communities a word is thought of as having (semantic, grammatical and phonological) unity and, in writing, words are conventionally separated by spaces. (In §9 below we investigate the writing convention when phonological and grammatical criteria do not produce the same unit.)

Indeed, in his *Phonemics*, Pike (1947: 89) defines 'word' as 'the smallest unit arrived at for some particular language as the most convenient type of

[7] The *up* can move to the left over an object that is a full NP but not over a preposition – *I made up the story* but not **I made up it*. Note the distinction between a phrasal verb like *make up* and one like *pick on*, where the *on* must precede the object NP, e.g. *He picked on his brother* or *He picked on him* but not **He picked his brother on* or **He picked him on*. See Dixon (1982; 1991: 274–8).

grammatical entity to separate by spaces; in general, it constitutes one of those units of a particular language which actually or potentially may be pronounced by itself'. Pike here implies that the ideal orthographic convention is to write spaces between grammatical words. The first part of his definition is circular – spaces are written around a grammatical word and a grammatical word is what is felt to be appropriately written between spaces; that is, no explicit criterion for 'grammatical word' is provided. The second part of his definition is essentially Bloomfield's 'minimum free form', discussed in §5 and §7 below.

Writing conventions are unlikely to be absolutely consistent. In English, for instance, the convention is to write *cannot* as one word but the analogous *must not* as two. There appears to be no reason for this; it is just a convention of the language community.

The Bantu languages of southern Africa have a complex but agglutinative verb structure. Van Wyk (1967: 230) describes different conventions used in these languages for writing word divisions:

(a) disjunctivism – 'according to which relatively simple, and, therefore, relatively short, linguistic units are written and regarded as words';
(b) conjunctivism – 'according to which simple units are joined to form long words with complex morphological structures'.

He exemplifies with the Northern Sotho sentence 'we shall skin it with his knife'. The two ways of writing this are:

(a) *re tlo e bua ka thipa ya gagwe*, according to the disjunctive system;
(b) *retloebua kathipa yagagwe*, according to the conjunctive system.

Van Wyk does not provide an interlinear gloss. However, we have been able to ascertain that *re-* is the 1pl subject prefix, *-tlo-* is the future prefix, *-e-* is a 3sg object prefix, *-bua* is the verb root 'to skin', *ka-* is an instrumental prefix, *thipa* is the noun 'knife', *ya-* is a class 9 prefix (agreeing with the class 9 noun 'knife') and *gagwe* is 'his'.

In fact different orthographic strategies have been adopted for different Bantu languages. Northern Sotho, Southern Sotho and Tswana are written disjunctively while Zulu and Xhosa are written conjunctively. There is no inherent grammatical difference between these languages; it is just that different writing conventions are followed. In the conjunctive system spaces are written between grammatical words (which may be long); in the disjunctive system spaces are written between morphemes within grammatical words. This may have been influenced by the fact that some of the prefixes are bound pronouns and case-type markers, corresponding to free pronouns and prepositions in languages such as English and Dutch (the languages of the Europeans who helped devise these writing systems), which are there written as separate words.

The orthographic conventions used for a language tend to reflect what the language was like at the time when an orthography was first adopted. For

example, *knee* was pronounced with an initial *k* when English was first written. A language may undergo considerable changes, few of which get incorporated into the orthography. French, for instance, has shifted from a mildly synthetic structure to one bordering on the polysynthetic. A sentence such as *je ne l'ai pas vu* 'I have not seen it' can be considered a single word, on both grammatical and phonological criteria. But the language is – as a reflection of its history – written disjunctively, with the consequence that speakers will say that the sentence consists of five or six words (see Vendryes 1925: 87–8). This is one of the reasons why linguists have found it harder to decide 'what is a word' for French than for many other languages. (This point is further pursued by Matthews in chapter 11.)

4.3 Grammatical and phonological aspects

Before the idea (followed here) that one should deal separately with 'grammatical word' and 'phonological word' and then examine the relationship between the two units, there was confusion about exactly what a word is.

As Ullmann (1957: 46) points out 'since the word is the central element of the language system, it is natural for it to face both ways: not only is it the chief subject matter of lexicology, but it is dependent on phonology for the analysis of its sound-structure, and on syntax for the delimitation of its status in more complex configurations'. But is 'word' primarily a grammatical unit, with some phonological properties; or is it primarily a phonological unit, with some grammatical properties; or is it equally a unit in grammar and in phonology? Ideas have varied.

The majority opinion has been that 'word' is primarily a unit of grammar although, as Matthews (1991: 209) notes 'the word tends to be a unit of phonology as well as grammar. In Latin, for example, it was the unit within which accents were determined'. Jespersen (1924: 92) states 'words are linguistic units, but they are not phonetic units' and Bloomfield (1933: 181) agrees that 'the word is not primarily a phonetic unit', while Meillet (1964: 136) maintains: 'le mot n'admet pas, comme la syllabe, une définition phonétique; en effet la notion de mot n'est pas phonétique, mais morphologique et syntaxique'.

Lyons (1968: 200–1) puts it this way: 'we will continue to assume, with the majority of linguists, that in all languages the morpheme is the minimum unit of grammatical analysis. The question we have set ourselves therefore is this: how shall we define a unit intermediate in rank between the morpheme and the sentence and one which will correspond fairly closely with our intuitive ideas of what is a 'word', these intuitive ideas being supported, in general, by the conventions of the orthographic tradition?' He then adds (p 204): 'in many languages the word is phonologically marked in some way'.

Pike (1947: 90) makes a clear distinction between 'grammatical units', which include 'morphemes, words, clitics, phrases and utterances', and 'phonological

units', which include 'phonemes, syllables, stress groups, rhythm groups, intonation groups, utterances, and so on'. Halliday puts forward a similar view, having 'word' as one of the five 'units' in his grammatical theory, the full list being: morpheme, word, group, clause and sentence (see, for example, Halliday, McIntosh and Strevens 1964: 25).

Just a few linguists opt for the opposite position. Newman (1967: 182–3) begins his perceptive study of words and word classes in Yokuts with lists of phonological and grammatical criteria, stating 'morphological criteria serve to supplement the phonological features for delimiting the unit word'. And Wells (1947: 99) states 'because of their insufficiency, the phonemic criteria of a word must be supplemented, for every or nearly every language, by criteria of the second kind . . . the grammatical'.

Utilising phonological and grammatical criteria to define a single unit can, not unnaturally, lead to conflicts and ambiguities. Wells rightly states – working in terms of a single unit 'word' – 'in fact, the word is most solid as a unit in those languages where phonemic and grammatical criteria reinforce each other'.

An alternative position is to provide a set of criteria for deciding 'what is a word' that mix grammatical and phonological features, with no indication of what should be given priority when they do not provide the same result; see, for example, Bazell (1957: 25–6) and Chao (1968), discussed in §5.

We will – in §6 and §7 – suggest definitions for phonological word and for grammatical word, which should in each instance give a clear and unambiguous result. We will also, in §8, briefly discuss clitics, which may constitute a grammatical word but not an independent phonological word (clitics are discussed more fully in chapter 2). Before that it will be instructive to look – in §5 – at some of the types of criteria that have been put forward in the literature.

5 Some suggested criteria

In a short but classic discussion of 'the word' Bazell (1953: 67–8) states that 'criteria may be found which are either necessary, or sufficient, but not both'. If criterion X is necessary but not sufficient for defining 'word' this implies that all words show X but some other units show X as well. If criterion X is sufficient but not necessary this implies that any unit showing X is a word but there are also some words that do not show X.

Bazell then provides examples: 'the vowel-congruence [vowel harmony] of alternating morphs is a sufficient but not necessary criterion of word-unity in Turkish; the presence of at least one vowel is a necessary but not a sufficient criterion of word-status in English. The possibility of pause is a sufficient criterion, in most languages, of word-division'.

Lyons (1968: 200) paraphrases Meillet: 'a word may be defined as the unit of a particular meaning with a particular complex of sounds capable of a particular