WORD, 2023
Vol. 69, No. 3, 283–297, https://doi.org/10.1080/00437956.2023.2237272

Defining the word

Martin Haspelmath *

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

In this paper, I propose a definition of the term *word* that can be applied to all languages using the same criteria. Roughly, a word is defined as a free morph or a clitic or a root plus affixes or a compound plus affixes. The paper relies on earlier definitions of the terms *free*, *morph*, *affix*, *clitic*, *root*, and *compound*, which are summarized here. I briefly compare the proposed definition with Bloomfield's, I note that it is a shared-core definition, and I say how *word-forms* differ from *lexemes*. In the final section, I explain why I think that an unnatural-seeming definition is better than a prototype definition or other options.

Keywords: Word; clitic; affix; compound; lexeme

1. The rationale

Since the early twentieth century, linguists have generally been aware that the traditional concepts of the European grammatical tradition may not carry over to languages in general (e.g., Boas 1911; Vendryes 1921). It has become increasingly clear that even a basic notion like 'word' may not be universal, or that in any event it cannot be taken for granted. For example, it is often contentious whether clitics such as English genitive '*s* are suffixes (see 1a), whether expressions such as Italian *carta di credito* are compound words (see 1b), and whether verbal markers such as German infinitival *zu* 'to' and particle *aus* 'out' are prefixes (see 1c). Most languages have such unclear cases and they cannot be dismissed as marginal.

(1)	a.	English	*Kim's garden*		
	b.	Italian	*carta di credito*	'credit card'	
	c.	German	*aus-zu-gehen*	'to go out'	(lit. 'out-to-go')

Increasingly over the last century, linguists have made use of new technical terms such as 'morpheme' and 'constituent' in order to talk about language structures in a more neutral way. But the term *word* never went away, and it is still true in linguists' practice that '*words* are the most basic of all linguistic units' (the first sentence in the *Oxford handbook of the word*, Taylor 2015). In highly technical papers in some subdisciplines, linguists may manage without using *word* as a term, but in practice, most of us talk

*Email: martin_haspelmath@eva.mpg.de

about words all the time. The very fact that we have separate textbooks for morphology (i.e., word structure) and syntax (i.e., word combinations) shows how important words are in our everyday work, and there are many areas of linguistics where the distinction between affixes, words, and phrases is of central theoretical importance (such as realizational morphology, e.g., Spencer 2013; grammatical complexity research, e.g., Lupyan and Dale 2010; phonological domains, e.g., Nespor and Vogel 2007; dependency syntax, e.g., Osborne and Gerdes 2019).

However, the literature so far contains no clear definition of *word* (either in the sense of 'word-form', or in the sense of 'lexeme'), and our textbooks or handbooks tend to gloss over this topic, or they deal with it by discussing a range of different criteria without coming to a clear conclusion. In Hippisley and Stump's 800-page *Cambridge handbook of morphology* (2016), only Nichols mentions the issue and lists a few criteria, but she concludes that 'no criteria are watertight' (Nichols 2016, 711). So it seems that we do not know how to identify words in general linguistics (Haspelmath 2011; Tallman 2020), and that linguists continue to operate with a vague prescientific concept that may be based on a stereotype derived from the orthographic conventions of European languages.

The solution to this problem cannot consist in avoiding the use of the term *word*, because this is not practical. So here I propose an alternative solution: In §2, I give a fairly simple technical definition with sharp boundaries that corresponds very largely to the intuitive notion of 'word' that most linguists (and lay people) have. The definition is technical in that it makes reference to other technical concepts that cannot be understood without some knowledge of general grammar. It is thus far more precise than Sapir's (1921, 34) definition ("the smallest, completely satisfying bits of isolated meaning into which the sentence resolves itself"), but at the same time not as complex and hard to understand as the definition provided by Mel'čuk (1993, 187–223), which takes dozens of pages to explicate. Then in §3, I deal with a number of questions that may arise from the definition, and in §4 I discuss a few alternative ways of addressing the lack of a commonly accepted definition. The focus of the present paper is on the definition of word in the sense of 'word-form' (i.e., text word) rather than 'lexeme' (i.e., dictionary word), but I will offer a definition of the latter term, too (§3.3).

It should be noted that the definition is proposed here as a contribution to the methodology of linguistics, and no substantive claims are involved. Some linguists think that our technical terms should define theoretical constructs (e.g., innate categories of universal grammar), but for most work in linguistics, this is not practical, because the proposals made within formal frameworks are generally quite tentative, and none are uncontroversial (see Stewart 2016 for a survey of diverse morphological frameworks or theories). By contrast, it is largely uncontroversial how we understand 'word', and by proposing a definition, I merely make this understanding precise in such a way that further work can build on it (or decide not to build on it).

2. The definition

The proposal of this paper is that 'word' (in the more specific sense of 'word-form' rather than 'lexeme') can be defined as a comparative concept for general linguistics as in Definition 1.

Definition 1: **word**
A word is (i) a free morph, or (ii) a clitic, or (iii) a root or a compound possibly augmented by nonrequired affixes and augmented by required affixes if there are any.

There are thus four kinds of forms that are word-forms, and a few English examples of each type are given in (1a–d).

(1)	a.	free morph	*nice, work, now, ouch*
	b.	clitic	*the, to, 's*
	c.	root (plus affixes)	*tree, nice-r, go-ing, re-work, re-place-ment-s*
	d.	compound (plus affixes)	*flower-pot, wind-shield, dog-sit, flower-pot-s*

I will say a bit more about the distinction between comparative-concept terms and descriptive categories of particular languages in §3.2. In the remainder of this section, I will give brief characterizations of the six technical concepts needed to understand the definition: *free form* (Definition 2), *morph* (Definition 3), *clitic* (Definition 4), *affix* (Definition 5), *root* (Definition 6), *compound* (Definition 7), and *required affix* (Definition 8).

We begin by defining the term *free form*, following Bloomfield (1933, 160) (see Haspelmath 2021a, §4).

Definition 2: **free form**
A free form is a form that can be used on its own.

Not only entire sentences can be used on their own (e.g., *Today my husband is working in a nice café*), but also nominal phrases (*my husband*) or adverbial phrases (*today, in a nice café*), especially as elliptical answers to questions (*who? when? where?*). But smaller parts of sentences are often not free (but bound) in that they cannot be used in isolation. In English, adnominal forms like *my* or *a*, and many verb forms such as *is* are bound forms which cannot occur on their own but must always cooccur with other forms (e.g., *my husband, a café, he is working*).

It has sometimes been thought that free vs. bound occurrence is a phonological property,[1] but this is not true. There is no phonological reason why the English adnominal modifier *my* is a bound form (cf. the noun *sky*) or why the verb form *is* should not be a free form (cf. the verb *fizz*). That they are bound forms is an idiosyncratic syntactic property: They simply have to cooccur with other forms.

Next, the term *morph* is defined as in Definition 3 (Haspelmath 2020a).

Definition 3: **morph**
A morph is a minimal form.

A form is a string of continuous segments that can be said to have a meaning, and a minimal form is a form that is not composed of several other forms. In the literature,

[1] E.g., Lyons (1968, 201): "Forms which never occur alone as whole utterances (in some normal situation) are *bound* forms; forms which may occur alone as uttences are *free* forms … It will be evident to the reader that this definition applies … to phonological words rather than grammatical words."

we more often find the older term *morpheme* used in roughly the same sense, for forms that are not composed of other forms, but *morpheme* has many different uses in linguistics (see Mugdan 2015). Commonly it is used in an abstract sense, e.g., for the set of all morphs that share the same meaning and syntactic behavior (so that, for example, the German plural suffixes *-e, -(e)n, -s* are said to 'belong to the same morpheme'), and morphemes can even be realized as zero. A morpheme is thus not necessarily a kind of form, but what exactly linguists mean by *morpheme* is often unclear (for quite different perspectives, see Anderson 2015 and Embick 2015). A morph is a concrete form that can be pronounced, though it may show phonological shape variations (for example, the English past-tense suffix *-ed* may be pronounced [d] or [t] or [əd], but it is a single suffix, i.e., a single morph).

When a morph (i.e., a minimal form) is a free form (i.e., when it can occur in isolation), it is a word, as in the examples in (1a), repeated here.

(2) nice, work, now, ouch

All these are not composed of other forms (and are thus morphs), and they can occur on their own (and are thus free forms). Interjections like *ouch* can be spontaneous utterances by themselves, and words like *nice*, *work* and *now* can be answers to questions (e.g., *what's the café like? what did he do? when should we talk?*).

Another kind of word is a clitic, defined as in Definition 4 (Haspelmath 2023a). This very simple definition presupposes the more complex definition of *affix* that is given below.

Definition 4: **clitic**
A clitic is a bound morph that is neither a root nor an affix.

Examples of English clitics were given in (1b): *the, to, 's.* They are not roots because they are not contentful forms (forms denoting objects, actions or properties; see Definition 6 below), and they are not affixes because they combine with roots of different classes.[2]

Next, we need a definition of *affix*, because one of the most important ways in which composite words differ from combinations of words is by combining a root with an affix. Definition 5 serves to delimit affixes from clitics (Haspelmath 2021a, §6).

Definition 5: **affix**
An affix is a bound morph that is not a root, that must occur on a root, and that cannot occur on roots of different root classes.

Like clitics, affixes are bound morphs that are not roots, but in addition, they must always occur on roots of the same class, i.e., always on verb roots, noun roots, or adjective roots. For example, the German infinitival marker *zu-* (in *aus-zu-gehen* 'to

[2] Clitics are often thought to be defined in phonological terms, as somehow 'phonologically dependent' elements, but there is no general agreement on what phonological dependence means. As is discussed extensively in Haspelmath (2023a), the phonological effects can be of diverse kinds and mostly apply to affixes as well.

go out' in (1c)) is a prefix (=a preposed affix) because it always occurs directly preceding a verb root (or a verb root preceded by another prefix). Unlike English infinitival *to*, it cannot precede an adverb (cf. examples like English *to thoroughly destroy*), so the English *to* is a proclitic, while the German *zu*- is a prefix. The German genitive -*s* is a suffix because it always follows a noun (e.g., *Kim-s Ring* 'Kim's ring'), but English genitive '*s* is an enclitic because it may follow not only nouns (as in *Kim*'*s umbrella, the dog*'*s bone*), but also verbs (e.g., *the boy I love*'*s umbrella*).[3]

The term *root* has been used several times already, and it is defined as in Definition 6:

Definition 6: **root**
A root is a contentful morph (i.e., a morph denoting an action, an object or a property) that can occur as part of a free form without another contentful morph.

This definition starts out from the generally agreed characterization of roots as content items (as opposed to function items),[4] following the definition given by Bauer et al. (2013: 17): "A root is the centre of a word, a lexically contentful morph, either free or bound, which is not further analysable." As I discuss in Haspelmath (2023b), an action root is much the same as a verb root in a comparative perspective, an object root is a noun root, and a property root is an adjective root (see also Croft 2000 on the semantic classes that are the basis of word classes in the world's languages).[5] Of course, some nouns denote abstract properties (e.g., *luck, soul, peace*), and these are not covered by the definition (see §3.2 below on shared-core definitions). The qualifying clause ('that can occur as part of a free form without another contentful morph') excludes morphs such as causative affixes (which denote actions of causation but always occur together with another root) or affixes with 'root-like' meaning such as Bella Coola -*ak* 'hand' and -*uc* 'mouth' which only occur together with a root (Mithun 1997). English may also be said to have such 'lexical affixes', e.g., *geo*- and *socio*- and other members of 'neoclassical compounds' that do not occur on their own.

Another classical delimitation issue concerns the term *compound*, because there is often a question whether expressions like Italian *carta di credito* 'credit card' should be included in the compound concept. My proposal is to exclude them and to limit compounds to constructions with no markers intervening between the two roots, as in Definition 7.

[3] Clitics are often said to combine with a phrase, but when a nonroot bound morph only ever occurs peripherally to one type of phrase on a content root, it is usually considered an affix (as is the case with complementizer suffixes on verbs in verb-final languages such as Turkish or Japanese, or case suffixes in languages with noun-final nominals such as Lezgian). Such elements are affixes by Definition 5.

[4] The term *root* is sometimes used in an abstract sense (e.g., for the triconsonantal skeletons in Semitic languages), but the definition used here refers to concrete forms.

[5] Needless to say, within each particular language, word classes are defined by morphosyntactic criteria. It is only at the level of comparative concepts that we need to appeal to semantic criteria.

Definition 7: **compound construction**
A compound construction is a construction consisting of two strictly adjacent slots for roots that cannot be expanded by full nominal, adjectival, or degree modifiers.

The first requirement is that the compound constituents must occur strictly next to each other, so that English *take part* is not a compound (because affixes can intervene as in *take-s part*, *tak-ing part*), and neither is German *Liebe-s-brief* 'love letter' or Italian *carta di credito*.[6] Excluding such cases is an arbitrary decision, of course, but it is in line with much current practice (where 'phrasal compounds' like *carta di credito* are often seen as types of multi-word constructions rather than complex words, cf. Masini 2019). The second requirement is that the roots cannot be expanded by nominal or adjectival modifiers. For example, in the compound *white-board*, the first compound member cannot be modified by an adjectival adverb (e.g., *brilliantly*), and in *flower-pot*, the first member cannot be modified by a possessor (e.g., *dad's*).

(3)	a.	*white-board*	**[brilliantly white]- board*
	b.	*flower-pot*	**[dad's flower]- pot*

It is often said that the modifying element in a compound must be nonreferential, but this applies only to nouns, and it is not always clear how to determine referentiality. The criterion of non-expandability by adjectival or nominal modifiers is simpler and easily applicable to any language.[7] For more on this definition of a compound, see Haspelmath (2024).

Finally, we need a notion of *required affix* because some roots and compound are words by themselves (e.g., English *tree, flower-pot*), while others are not and must be combined with an affix. For example, Italian *alber-o* 'tree' must include the singular suffix *-o*, and German *geb-en* '(to) give' must include the infinitive suffix *-en*.

Definition 8: **required affix**
A required affix is an affix that must be present in a free form unless it is replaced by another affix.

The singular suffix *-o* must be present in *alber-o* because the root *alber* 'tree' cannot occur in a free form, but it may be replaced by the plural suffix *-i* (*alber-i* 'trees'). In German, the verb root *geb* 'give' must occur with the suffix *-en* (*geb-en* 'to give'), or it may take an alternative required suffix (e.g., *geb-t* 'you (PL) give (imperative)'). By contrast, English *tree* has no required affix: The plural suffix *-s* is not required because it can be absent and need not be replaced by another affix in a free form such as *a tree.*

[6] This requirement is violated even more clearly in German *teil-nehmen* 'take part', because the element *teil* sometimes occurs postverbally (cf. *sie nehmen morgen teil* 'they will take part tomorrow', vs. *sie haben gestern teil-genommen* 'they took part yesterday'). The spelling is not in conformity here with the compound status according to Definition 7.

[7] In many Indo-European languages, compounds can consist of a combination of a root and a compound, which is not reflected in Definition 7. For the sake of simplicity, this is not addressed in the present paper. Moreover, parasynthetics (of the type 'root+root+affix', e.g., *blue-eye-d*) are not included either.

So if there are any required affixes, the root (or compound) is not a word, but if there are none, the root (or compound) is a word. For example, English *tree* is a root and *flower-pot* is a compound, and while they are not free forms (because they are singular count nouns and must be combined with an article in English), they are words according to Definition 1 because they do not have required affixes (the plural suffix *-s* is not a required because it can be absent without being replaced by another affix). Composite forms consisting of a root plus further affixes (e.g., *re-place-ment*) are words too, again unless there are any required affixes.

3. Some questions that may arise

3.1. *How does the definition relate to Bloomfield's (1933)?*

There is only one reasonably short definition of *word* that has become widely known: Bloomfield's (1926, 1933) definition of a word as a 'minimal free form':

> A minimum free form is a word. A word is thus a form which may be uttered alone (with meaning) but cannot be analyzed into parts that may (all of them) be uttered alone (with meaning). (Bloomfield 1926, 156)

This definition has often been cited and adopted (e.g., by Hockett 1958, 168; Kiparsky 2020), but it is both too narrow and too broad. On the one hand, it is too narrow because it does not include clitics, which are bound forms but which are generally regarded as words. Moreover, it does not include nouns which must be accompanied by an article, such as English *tree*:

(9) a. *tree* (cannot be used on its own)
b. *a tree* (can be used in isolation)

By contrast, mass nouns such as English *work* or *fire* can be used in isolation. In Definition 1, English count nouns such as *tree* are included because they are roots with no required affixes, even though they are bound.

On the other hand, Bloomfield's definition is too broad because it includes some combinations of clitics and free forms, e.g., English *a tree,* or *to Leipzig*, or *put it there.* These are free forms, and they are all minimal in the sense that they cannot be broken up into parts that are free forms. In *a tree*, neither of the parts is a free form; in *to Leipzig*, the preposition is not a free form; and in *put it there*, neither *put* nor *it* are free forms (nor is *put it*).

Bloomfield (1933, 179) actually recognized the latter problem, but he did not offer a general solution. And the subsequent literature has not tried to provide a definition that builds on, but corrects the defects of, Bloomfield's definition.

3.2. *Why are the definitions not always 'exact'? On shared-core definitions*

Some of the definitions of §2 do not correspond exactly to the definitions of the corresponding language-particular (or descriptive) terms when applied to specific languages. This is perhaps clearest in the cases of *compound* and *clitic*. A German Nominal Compound (as a language-particular category) is defined with respect to stress and inflectional properties, so *Rotwein* (in 10a) is a compound (stress on the first part, no inflection on the adjective), while *Weißes Haus* (in 10b) is not a compound (primary stress on *Haus*, inflection on the adjective *Weiß-es*).

(10) German

a.	*Rótwein*	'red wine'
b.	*Wèißes Háus*	'White House'

But stress and inflection are not cross-linguistically general properties, so a general (comparative-concept) definition of *compound* cannot make reference to them (see Definition 7). Similarly, *clitic* might be defined in language-particular terms with respect to stress (e.g., in Ancient Greek, where Clitics are defined as stressless bound words that may influence the stress of the preceding word), but from a cross-linguistic perspective (as a comparative concept), the term needs to be defined without reference to prosodic properties (see Definition 4). In general, applying general definitions to specific languages leads to some apparent 'inexactness', as when it turns out that German *Liebe-s-brief* 'love letter' does not count as a compound, though many other cases do (e.g., *Auto-bahn* lit. 'car-way', *Flug-hafen* lit. 'flight-port').

But this does not mean that there is any kind of vagueness or imprecision in the definitions. The reason for the (possibly surprising) discrepancies is that languages differ in a wide variety of ways and comparative concepts rarely capture their similarities fully. Definitions of commonly used grammatical terms are typically SHARED-CORE DEFINITIONS, in that they capture what is shared among certain similar categories, but not the precise language-particular boundaries (see Haspelmath 2021b, §5 for this concept). What German and English compounds (and compounds of all other languages) share is captured by Definition 7, but the corresponding language particular categories (German Compounds and English Compounds) are defined in language-particular terms.[8]

3.3. *What is a lexeme as opposed to a word-form?*

The definition that I provided in §2 is a definition of *word* in the sense of 'word-form', i.e., a word as it may be used in a text.[9] But we also talk about dictionary words, which are often called *lexemes*. This term has become better known since the 1960s (e.g., Lyons 1968; Matthews 1974) and is now fully established, but what exactly is a lexeme? I discuss this in Haspelmath (2023c), and here I merely provide a brief summary. The main point is that a lexeme is not a kind of form, but a set of word-forms. I propose that it should be defined on the basis of the notion of *lexeme-*

[8] For the spelling difference (lower-case *compound* as a comparative concept; upper case *German Compound* for the language particular category), see Haspelmath (2020b, §3).

[9] Other terms are *grammatical word* (Matthews 1974, 32) and *morphosyntactic word* (Haspelmath 2011, 38). The term *word-form* comes from the Russian tradition (e.g., Mel'čuk 2006). But note that *grammatical word* has been used to differentiate between homophonous word-forms (e.g., English Past Tense *play-ed* vs. Past Participle *play-ed*, which are the same word-form according to Gebhardt (2023, 83), but two different grammatical words). This usage is non-standard.

stem, which itself is based on the notion of *inflectional affix*. Informally, we can say that a lexeme-stem is something that we get when we strip an inflected form of its inflectional affixes. A few simple examples are given in (11). German *Auto-bahn* [car-way] and Spanish *juga-dor* [play-AGENT] show that lexeme-stems may contain several roots or derivational affixes, but they do not contain inflectional affixes.

(11)	**language**	**lexeme**	**lexeme-stem**	**some word-forms in the set**
	English	WALK	*walk-*	*walk-s, walk-ed*
	Latin	LUPUS 'wolf'	*lup-*	*lup-us* 'NOM.SG', *lup-i* 'NOM.PL'
	German	AUTOBAHN 'freeway'	*Auto-bahn-*	*Autobahn-en* 'freeways'
	Spanish	JUGADOR 'player'	*juga-dor-*	*jugador-es* 'players'

Then we can define *lexeme-stem* as in (12). Note that the distinction between inflectional and derivational affixes is often thought to be problematic, but in Haspelmath (2023c), I define them (somewhat arbitrarily) in semantic terms.

(12) **lexeme-stem**:
A lexeme-stem is a root, or a compound, possibly augmented by derivational affixes, which can combine with inflectional affixes if the language has any but does not contain any inflectional affixes.

Linguists sometimes treat lexemes as if they were forms, not abstract entities,[10] and when they do this, they seem to have something like the notion of lexeme-stem in mind. Informally, one can of course use *lexeme* in the sense of 'lexeme-stem', but it should be kept in mind that the more correct term for a kind of form that does not include (but could be combined with) inflectional affixes is *lexeme-stem*.[11]

On the basis of this notion of lexeme-stem, we can define a lexeme as in (13).

(13) **lexeme**:
A lexeme is the set of forms that minimally contain the same lexeme-stem, or one of its suppletive counterparts, and that may only contain inflectional affixes in addition.

As this definition is based on a fairly arbitrary delimitation of inflectional from derivational affixes (Haspelmath 2023c), the resulting notion is not particularly natural, but I wanted to give it here anyway for completeness.

[10] For example, Breiter (1994) investigates 'the length of lexemes in Chinese', and what she means is the length of the lexeme-stem. As there are few inflectional affixes in Chinese, the difference does not matter much, but it is important to be aware that only lexeme-stems are forms (with a measurable length), while lexemes are sets of forms.

[11] Occasionally, inflectional affixes occur directly on the root and derivational affixes occur outside of them (e.g., Ancient Greek *an-é-bain-on* [up-PST-go-1SG] 'I went up', where the inflectional past tense prefix occurs closer to the root *bain-* 'go' than the derivational prefix *an(a)-*). Such cases are not covered by this definition, which concentrates on the core phenomena and thus follows the principle that comparative concepts should be 'shared-core definitions' (§3.2).

4. Confronting the absence of a definition of *word*

I have provided a definition of *word* (§2) and have discussed some further questions that readers may have (§3), but now I want to put the present proposal into a larger context as a way of justifying my proposal further. After all, linguists have been aware of the lack of a unique and clear definition of *word* for quite some time. According to Alpatov (2018, 21), Aničkov (1963) collected 34 different definitions of *word* (Russian *slovo*) found in works of the nineteenth and twentieth century. None of them (except for Bloomfield's) have become widely known, so I will assume that it is still true that 'in general, there is no satisfactory definition of the word' (Žirmunskij 1966, 65). In this section, I distinguish five ways of confronting this problem, and in the final subsection, I explain why I opt for the fifth approach (§4.5).

4.1. *The definition of word as a philosophical-historical question*

Some authors appear to treat the word *word* like other highly general concepts, such as 'freedom', or 'happiness', or 'matter', i.e., as a word or concept whose significance is taken for granted, though there are many different ways in which it can be conceptualized or integrated into a larger view of the world. Comprehensive surveys such as Mugdan (2015, §3) often leave the reader wondering about the ultimate purpose of such an exercise. We all know that different scientists in the past used diverse concepts and methods and came to diverse conclusions, but how do modern readers profit from surveys that have no conclusions?[12] If each generation of linguists merely adds a few further considerations but all the earlier ideas remain relevant because the controversies have not been resolved, then linguistics is treated as a branch of philosophy, where modern scholars still discuss Aristotle and Kant as their equals.[13]

But this is not how I want to see linguistics, which I regard as an empirical science like biology. I therefore think that we should either not use the term *word* (§4.2) or provide a definition (§4.5).

4.2. *Avoiding the term word*

Perhaps the most straightforward conclusion that could be drawn from the lack of a generally accepted definition of *word* is that one should refrain from using the term. After all, some older expressions have fallen out of use: When linguists realized that the differences between cognate words (such as English *pound* and German *Pfund*) concerned the sounds that speakers made, they stopped speaking about 'letters'. Instead, they started using terms like *sound*, or *phoneme*, or *distinctive segment*. So instead of *word*, we could use more technical terms, such as *root*, *affix*, *verb complex*. This was the view I adopted after realizing (in Haspelmath 2011) that linguists had no generally applicable definition of *word*. I also found similar problems

[12] Mugdan (2015, 252) writes at the end: "Where there is room for legitimate disagreement … , the final decision should be guided by overall structural considerations but may often well be a matter of taste."

[13] Ryan Nefdt (p.c.) pointed out to me that there is a tradition in philosophy of thinking about the ontology of words (e.g., Miller 2020). It remains to be seen how that work relates to the work in linguistics that is in my purview here.

with the term *clitic* (Haspelmath 2015) and proposed that it should likewise be abandoned.

However, as I noted earlier in §1, avoiding *word* is not practical as a general solution for linguistics, because the term is too deeply entrenched in the field. It is not like older scientific terms such as *phlogiston* (an old term from chemistry) or *aether* (in physics), which were proposed by scientists in the context of specific theories which were later superseded by new theories. Linguistics has specific theories in which morphology and syntax are distinct modules and there is thus a clear place for a technical term *word* (e.g., Anderson 1992; Bresnan and Mchombo 1995), but the term is widely used and understood throughout the discipline, far beyond these specific approaches. As I noted, even the difference between morphology and syntax is based on the 'word' notion. It is thus not an option to hope that the term *word* will disappear the way *phlogiston* and *aether* disappeared. In languages with orthographic word separation, words are simply too salient, so that linguists will keep using it.[14]

4.3. *Ignoring the problems with the definition*

Another approach one might take is to pretend that all is well and that there are no problems with the definition of *word*. This is what many linguists have been doing implicitly all along, but it is particularly striking in Aikhenvald, Dixon, and White (2020), an update of the same authors' treatment of the issue in 2002 (Dixon and Aikhenvald 2002). They still appeal to the vague notion of 'conventionalized coherence and meaning' (even though it was criticized in Haspelmath 2011), and they do not provide clear criteria for distinguishing between clitics and affixes, and between phrases and compounds. If one is already convinced that "'word' is a pivot for every language" (Aikhenvald, Dixon, and White 2020, 1), then it may not be a big problem that there is no clear definition. But this certainly cannot be a general solution for the field of linguistics.

4.4. *The word as a prototype*

Over the last few decades, a common view among linguists has been that some traditional grammatical concepts may not have clear boundaries, but they are cross-linguistic prototypes. A clear statement of this view is found in Taylor (1995, §10.1), who notes that a general definition of *word* has been difficult, but that languages have 'intuitively clear cases alongside a number of not-so-clear cases', which suggests that *word* is a prototype category. After discussing a range of different criteria for *word*, *affix* and *clitic*, he concludes:

> "There are good, representative examples of words *(mother),* of affixes *(-ed),* and of clitics (Zulu *ke);* there are not such good examples *(the);* and there are borderline cases." (Taylor 1995, 181)

[14] However, the salience of the term *lexeme* has been reduced in recent times, as we no longer need lexeme-based dictionaries to look up information about words. Modern technology provides us easy ways to search for meanings and other properties of words that do not involve the notion of lexeme (which seems to be rooted in the dictionary word; see Haspelmath 2023c).

It is my impression that this view of the nature of general linguistic categories is quite widespread and is typically adopted implicitly by linguists. The existence of borderline cases is often accepted as a fact of life but is rarely seen as putting the primary distinctions in question.

However, if we do not have independent evidence for the existence of these general categories, then the prototype view is not empirically responsible. It is an open question whether the phenomena that we find in languages group together in such a way that words, affixes and clitics are real clusters in the data. If these three terms merely stand for traditional stereotypes derived from European languages, then nothing is gained by saying that non-fitting cases are 'intermediate' or 'non-prototypical' (see also Tallman and Auderset 2022).

4.5. *Providing an unnatural definition*

The final approach is the one taken in this paper: Instead of failing to provide a definition (§4.1-3) or resorting to a vague prototype definition (§4.4), I propose that linguists are best served by a clear definition even if it looks quite unnatural, like Definition 1 above. Note that not only the definition of *word* is unnatural in the sense that it is complex and disjunctive, but that this applies also to the definitions of *affix*, *clitic*, *compound* and *lexeme* that we saw in §2 above.

By the informal expression 'unnatural definition', I mean a definition that does not seem to single out a unitary phenomenon, and that does not seem to *cut nature at its joints*. While a *morph* is a natural-looking entity (a minimal form, perhaps analogous to other minimal entities in science such as atoms and genes), words, affixes, clitics and compounds are defined in such a way that it would be very surprising if they were basic entities of general grammar.

The claim of this paper is that these definitions correspond reasonably well to the way in which these terms are actually used in linguistics. I hope that they are simple enough to be used in textbooks, general enough to make them applicable to all languages, and at the same time precise enough to cover at least 80% of the earlier usage.

5. Conclusion: on the uses of an unnatural definition

I conclude that the term *word* is best defined as in Definition 1, repeated here.

Definition 1: **word**
A word is (i) a free morph, or (ii) a clitic, or (iii) a root or a compound possibly augmented by nonrequired affixes and augmented by required affixes if there are any.

As noted in §4.5, this is not a natural definition, so why would I propose it here? The reason is that there is a widespread implicit view in linguistics that the unit *word* is a natural way of dividing up texts in languages. Moreover, there is a widespread view (implicit or explicit) that grammar is naturally divided into morphology and syntax, two modules that are perhaps linked by an 'interface', and that obey different regularities (e.g., lexical integrity as a unique property of morphology; Bresnan and Mchombo 1995).

However, it is quite possible that the salient general notion of *word* is an artifact of European spelling conventions, and that the idea of a morphology–syntax division derives from the tradition of European and North American linguistics which places great importance on the *word*. In other words, it could be that the notions of ‘word’, ‘morphology’ and ‘syntax’ are stereotypes derived from our tradition that have no actual basis in the reality of languages.

How will we find out whether words are real, in the sense that they are part of the nature of languages, and not just derived from our spelling-based preconceptions? One way in which we will definitely NOT find out is by simply ASSUMING that words are a crucial concept – as has been done by most syntax and morphology textbooks of the last few decades, and by much of the research literature.

Thus, I hope that the clear but complex (and unnatural) definition of *word* that I have offered here will stimulate linguists to carry out more research on the fundamental units of grammar, and that it will lead to a more cautious attitude toward traditional concepts which may be no more than inherited stereotypes.

Acknowledgements

For helpful comments on an earlier version of this paper, I thank Artemij Keidan, Susanne Michaelis, Ryan M. Nefdt, Andrey Shluinsky, Nina Speransky, Adam Tallman, Dmitrij Zelenskij, and further commentators on Academia.edu.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Martin Haspelmath http://orcid.org/0000-0003-2100-8493

References

Aikhenvald, Alexandra Y., Robert M. W. Dixon, and Nathan M. White. 2020. “The Essence of ‘Word’.” In *Phonological Word and Grammatical Word: A Cross-Linguistic Typology*, edited by Alexandra Y. Aikhenvald, Robert M. W. Dixon, and Nathan M. White, 1–24. Oxford: Oxford University Press.

Alpatov, Vladimir. M. 2018. *Slovo i časti reči* [The word and the parts of speech]. Moscow: LRC.

Anderson, Stephen R. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Anderson, Stephen R. 2015. “The Morpheme: Its Nature and Use.” In *The Oxford Handbook of Inflection*, edited by Matthew Baerman, 11–33. Oxford: Oxford University Press.

Aničkov, I. E. 1963. “Ob opredelenii slova” [On the definition of the word]. In *Morfologičeskaja struktura slova v jazykax različnyx tipov*, edited by Viktor M. Žirmunskij and O. P. Sunik, 146–159. Leningrad: Izdatel’stvo Akademii Nauk SSSR. https://doi.org/10.5281/zenodo.8180316

Bauer, Laurie, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Bloomfield, Leonard. 1926. “A Set of Postulates for the Science of Language.” *Language* 2 (3): 153–164. https://doi.org/10.2307/408741.

Bloomfield, Leonard. 1933. *Language*. New York: H. Holt and Company.

Boas, Franz. 1911. “Introduction.” In *Handbook of American Indian Languages*, edited by Franz Boas, 1–83. Washington, DC: Bureau of American Ethnology.

Breiter, Maria A. 1994. “Length of Chinese Words in Relation to Their Other Systemic Features.” *Journal of Quantitative Linguistics* 1 (3): 224–231. https://doi.org/10.1080/09296179408590020.

Bresnan, Joan and Sam A. Mchombo. 1995. “The Lexical Integrity Principle: Evidence from Bantu.” *Natural Language and Linguistic Theory* 13 (2): 181–254. https://doi.org/10.1007/BF00992782.

Croft, William. 2000. “Parts of Speech as Language Universals and as Language-Particular Categories.” In *Approaches to the Typology of Word Classes*, edited by Petra M. Vogel and Bernard Comrie, 65–102. Berlin: Mouton de Gruyter.

Dixon, Robert M. W. and Alexandra Y. Aikhenvald. 2002. “Word: A Typological Framework.” In *Word: A Cross-Linguistic Typology*, edited by Robert M. W. Dixon and Alexandra Y Aikhenvald, 1–41. Cambridge: Cambridge University Press.

Embick, David. 2015. *The Morpheme: A Theoretical Introduction*. Berlin: De Gruyter Mouton.

Gebhardt, Lewis. 2023. *The Study of Words: An Introduction*. London: Routledge. https://doi.org/10.4324/9781003030188.

Haspelmath, Martin. 2011. “The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax.” *Folia Linguistica* 51 (1): 31–80. https://doi.org/10.1515/flin-2017-1005.

Haspelmath, Martin. 2015. “Defining vs. Diagnosing Linguistic Categories: A Case Study of Clitic Phenomena.” In *How Categorical are Categories? New Approaches to the Old Questions of Noun, Verb, and Adjective*, edited by Joanna Błaszczak, Dorota Klimek-Jankowska, and Krzysztof Migdalski, 273–304. Berlin: De Gruyter Mouton.

Haspelmath, Martin. 2020a. “The Morph as a Minimal Linguistic Form.” *Morphology* 30 (2): 117–134. https://doi.org/10.1007/s11525-020-09355-5.

Haspelmath, Martin. 2020b. “The Structural Uniqueness of Languages and the Value of Comparison for Language Description” *Asian Languages and Linguistics* 1 (2): 346–366. https://doi.org/10.1075/alal.20032.has.

Haspelmath, Martin. 2021a. “Bound Forms, Welded Forms, and Affixes: Basic Concepts for Morphological Comparison.” *Voprosy Jazykoznanija* 2021 (1): 7–28. https://doi.org/10.31857/0373-658X.2021.1.7-28.

Haspelmath, Martin. 2021b. “Towards Standardization of Morphosyntactic Terminology for General Linguistics.” In *Linguistic Categories, Language Description and Linguistic Typology*, edited by Luca Alfieri, Giorgio Francesco Arcodia, and Paolo Ramat, 35–57. Amsterdam: Benjamins. https://doi.org/10.1075/tsl.132.02has.

Haspelmath, Martin. 2023a. “Types of Clitics in the World’s Languages.” *Linguistic Typology at the Crossroads* (to appear).

Haspelmath, Martin. 2023b. “Word Class Universals and Language-Particular Analysis.” In *Oxford Handbook of Word Classes*, edited by Eva van Lier, 15–40. Oxford: Oxford University Press. (to appear).

Haspelmath, Martin. 2023c. “Inflection and Derivation as Traditional Comparative Concepts.” *Linguistics.* (to appear).

Haspelmath, Martin. 2024. “Compound and Incorporation Constructions as Combinations of Unexpandable Roots.” (to appear). https://doi.org/10.5281/zenodo.8137251.

Hippisley, Andrew and Gregory Stump, eds. 2016. *The Cambridge Handbook of Morphology.* Cambridge: Cambridge University Press.

Hockett, Charles F. 1958. *A Course in Modern Linguistics.* New York: MacMillan.

Kiparsky, Paul. 2020. “Morphological Units: Stems.” In *Oxford Research Encyclopedia of Linguistics 2020*. https://doi.org/10.1093/acrefore/9780199384655.013.542.

Lupyan, Gary and Rick Dale. 2010. “Language Structure is Partly Determined by Social Structure.” *PLoS ONE* 5 (1): e8559. https://doi.org/10.1371/journal.pone.0008559.

Lyons, John. 1968. *Introduction to Theoretical Linguistics.* Cambridge: Cambridge University Press.

Masini, Francesca. 2019. “Multi-word Expressions and Morphology.” In *Oxford Research Encyclopedia of Linguistics 2019*. https://doi.org/10.1093/acrefore/9780199384655.013.611.

Matthews, Peter H. 1974. *Morphology.* Cambridge: Cambridge University Press.

Mel’čuk, Igor. 1993. *Cours de morphologie générale (théorique et descriptive): Volume 1*. Montréal: Presses de l’Université de Montréal.

Mel’čuk, Igor A. 2006. *Aspects of the Theory of Morphology*. Berlin: De Gruyter. https://doi.org/10.1515/9783110199864.

Miller, James T. M. 2020. “The Ontology of Words: Realism, Nominalism, and Eliminativism.” *Philosophy Compass* 15 (7): e12691. https://doi.org/10.1111/phc3.12691.

Mithun, Marianne. 1997. “Lexical Affixes and Morphological Typology.” In *Essays on Language Function and Language Type: Dedicated to T. Givón*, edited by Joan L. Bybee, John Haiman, and Sandra A. Thompson, 357–371. Amsterdam: Benjamins.

Mugdan, Joachim. 2015. “Units of Word-Formation.” In Vol. 1 of *Word-Formation: An International Handbook of the Languages of Europe*, edited by Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, 235–301. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110246254-017.

Nespor, Marina and Irene Vogel. 2007. *Prosodic Phonology*. 2nd ed. Berlin: Mouton de Gruyter.

Nichols, Johanna. 2016. “Morphology in Typology.” In *The Cambridge Handbook of Morphology*, edited by Andrew Hippisley and Gregory Stump, 710–742. Cambridge: Cambridge University Press.

Osborne, Timothy and Kim Gerdes. 2019. “The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD).” *Glossa: A Journal of General Linguistics* 4 (1): 17. https://doi.org/10.5334/gjgl.537.

Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & Co.

Spencer, Andrew. 2013. *Lexical Relatedness*. Oxford: Oxford University Press.

Stewart, Thomas W. 2016. *Contemporary Morphological Theories: A User’s Guide*. Edinburgh: Edinburgh University Press.

Tallman, Adam J. R. 2020. “Beyond Grammatical and Phonological Words.” *Language and Linguistics Compass* 14 (2): e12364. https://doi.org/10.1111/lnc3.12364.

Tallman, Adam J. R. and Sandra Auderset. 2022. “Measuring and Assessing Indeterminacy and Variation in the Morphology-Syntax Distinction.” In *Linguistic Typology*. De Gruyter Mouton. https://doi.org/10.1515/lingty-2021-0041.

Taylor, John R. 1995. *Linguistic Categorization: Prototypes in Linguistic Theory*. 2nd ed. Oxford: Clarendon Press.

Taylor, John R. 2015. “Introduction.” In *The Oxford Handbook of the Word*, edited by John R. Taylor, 1–20. Oxford: Oxford University Press.

Vendryes, Joseph. 1921. *Le Langage*. Paris: Renaissance du livre.

Žirmunskij, Viktor M. 1966. “The Word and its Boundaries.” *Linguistics* 4 (27): 65–91. https://doi.org/10.1515/ling.1966.4.27.65.