What is a word?

John D. Phillips

When the editors of this journal require that papers submitted in English should not exceed 4,000 words, the meaning is fairly clear. Words are separated by spaces so it is a simple matter to count them, with a few minor exceptions (*ice cream* is two words, *icecream* one word, but what about *ice-cream*? *I* is a word, as is *have*, but what about *I've*?) On the other hand, the size limit for papers in Japanese or Korean is 20,000 letters — counted in letters rather than words. Is this because these languages have no unit corresponding to English word? A glance at the original Japanese of the length instruction

> 論文を執筆する場合、その長さは、原則として和文の場合、400字詰め原稿用紙換算で、50枚（20,000字）以内、欧文の場合は4,000語以内、ハングルは20,000字以内、中国語は10,000字以内とする。

shows us that Japanese does not mark words off with spaces. Spoken English is, on the face of it, like Japanese: the spaces between written words are not pronounced. To the average English-speaker, words are the basic building blocks of language: speaking consists of putting words together; but would an illiterate English speaker divide the language up in the same way that standard spelling does, or are words just an artefact of the writing system?

English-Japanese dictionaries usually give 単語 or 語 as the Japanese translation of word, and indeed 語 is used in the Japanese quote above. However, monolingual Japanese dictionaries typically define the meaning of 単語 as the smallest building block of language representing a unit of meaning and having a grammatical function[1]. This sounds more like the definition of the linguistic term morpheme[2].

Dixon & Aikhenvald note that "only some languages actually have a lexeme[3] with the meaning 'word'. Even in some familiar languages where this does occur it may be a recent development."[4] Reamer goes further: "Just as languages often lack a precise equivalent for the term 'word', they also lack a precise equivalent for the English term 'meaning', in the sense of a stable linguistic property of words."[5]

The modern English convention whereby a word is that which is delimited by space or punctuation in writing, is clearly inconsistent, as the following examples show:

Nouns:	*One word*	*Two words*	***Others:***	*One word*	*Two words*
	skyscraper	sky writing		cannot	will not
	bookseller	book dealer		another	the other

[1] E.g.「文法上で、意味・職能を有する最小の言語単位。」小学読本（1873）「文法上、意味・職能をもった最小の言語単位。」小学館デジタル大辞泉（2023）

[2] Morpheme: the smallest meaningful unit of a language. E.g. the word *smallest* contains two morphemes, the root *small* and the superlative suffix *est; meaningful* contains three morphemes, *mean+ing+ful*.

[3] Lexeme: an item of vocabulary.

[4] R. M. W. Dixon and Alexandra Y. Aikhenvald: 'Word: a typological framework', p. 2. In: Dixon & Aikhenvald (eds.) , 2002, pp. 1-38.

[5] Nick Reamer, p. 305. In *Handbook*, pp. 305-319.

countryman	country person	into	out of
screwdriver	bus driver		
airliner	ocean liner		
wallpaper	art paper		
suitcase	pencil case		

Names like *New York* and *New Guinea* seem to be single words: we can say *a new green shirt*, but *New green York* is impossible, it has to be *green New York*, the two parts of the name cannot be split.

Some punctuation marks can both join and separate words:

Hyphenated words	*One word*	*Two words*
	non-alcoholic beer	Japan's best-known city
	non-commital	oil-fired heating
	the pre-Victorian period	a six-inch ruler
		a four-part song

Here *non-*, *pre-*, and *-commital* are not used as independent words. The apostrophe in *I'm*, or *five o'clock*, or *O'Reilley*, or *John's hat*, or *the Emperor of Japan's hat* also has an indeterminate function: *I'm* seems to be two words, but *o'clock* and *O'Reilley* might each be taken to be one word; *John's* could be one word, but in *the Emperor of Japan's hat* the hat is the Emperor's, not Japan's! This might seem to be a minor problem, but in a sample of about half a million words of English text, there was a difference of 11.5% between the number of words counted as separated by space, and the number of words counted as separated by space or punctuation.

The treatment of compound words varies from language to language. For instance the German word *Dampfschiffahrtsgesellschaft* can be analysed as

Dampf +	schiff +	fahrt +	gesellschaft
steam	*ship*	*travel*	*company*

German typically writes compounds as a single word where English usually writes the elements separately, particularly when there are more than two. This seems to be a matter of tradition rather than any linguistic difference.

Other languages which delimit words orthographically have similar problems. Of course many languages do not have the concept of orthographic word at all, including besides Japanese most other east and south-east Asian languages: Korean, Chinese, Tibetan, Javanese, Thai, Burmese, etc. Most of the world's languages are unwritten: with these orthography is simply irrelevant.

The term *word* has been applied to various, often incompatible, types of linguistic unit. A paper by Trask helpfully lists and discusses many of these uses[6]. Besides the orthographic word, the lexeme has been mentioned above. The lexeme is an abstraction which may have several wordforms, also often referred to simply as *words*. One wordform will be a citation form. An English verb for instance can have up to six different wordforms, e.g.

[6] L. Trask: *What is a Word?* University of Sussex Working Papers in Linguistics and English Language, 2004.

drink, drinks, drinking, drank, drunk, drunken are forms of the lexeme the citation form of which is *drink*. Most verbs have only four forms: additional forms of the lexeme cited as *walk* are *walks, walking, walked*. In some languages a single lexeme can have dozens or even hundreds of wordforms.

Phonological Criteria

Another type of word is the phonological word: “a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language.”[7] Some languages have rather clear criteria from vowel harmony or from prosody or by restrictions on the shapes of word-endings or beginnings.

In languages with vowel or consonant harmony, the word can be thought of as the unit within which the harmony operates. Finnish has vowel harmony so that affixes, as opposed to neighbouring words, vary their vowels to harmonise with those of the stem they attach to. All the vowels in a word must belong to the same vowel class, so the suffixes *-sta*, *-vat*, and *-ssa* appear with a front vowel [æ] when they are part of a word with front vowels, as on the left below (in IPA), but with a back vowel [ɑ] when they are part of a word with back vowels, as on the right below.

tyhmæ•stæ	stupidly	tuhmɑ•stɑ	badly
syø•væt	they eat	lɑulɑ•vɑt	they sing
pæivæ•ssæ	in a day	suome•ssɑ	in Finland

Prosody often shows word boundaries. In Japanese the lexical pitch accent can serve a similar delimiting function to vowel harmony as described above. A Japanese noun, for instance, has a distinctive pattern of pitches distributed over its syllables (actually morae). In Standard Japanese the accent can be described in terms of two pitches, high and low. Importantly, the noun’s pitch pattern extends to its case particle. For instance there are three Japanese lexemes with citation form pronounced はし *hasi*, each with a different accent pattern, meaning ‘chopstick’, ‘bridge’, and ‘edge’. The citation form for ‘chopstick’ has high-low accent, those for ‘bridge’, and ‘edge’ low-high:

spelling	箸	橋	端
translation	chopstick	bridge	edge
pronunciation	hasi	hasi	hasi
accent	H L	L H	L H

If we add any case particle, here が *ga* marking grammatical subject, we get three different pronunciations:

pronunciation	hasi ga	hasi ga	hasi ga
accent	H L L	L H L	L H H

If accent can be taken to delimit words, citation form plus case marker form a single word. The same argument can be made for verbs: tense and other particles following verbs are included in the verb’s accent. The

[7] Trask, *op. cit.*, §2.

pronunciation かった *katta* serves as the past tense of the verbs 買った 'bought' and 勝った 'won', but with two different pitch patterns, showing that the past-tense particle た *ta* is part of the phonological word. Nerida Jarkey has a recent English-language discussion of this data[8].

Welsh is another language where words are clearly delimited by prosody. There is a pitch accent on the final syllable of a word, and a stress accent usually on the penultimate syllable[9].

llong	llongwr	llongwriaeth	llongwriaethol
ˈɬóŋ	ˈɬoŋúr	ɬoŋˈurjáeθ	ɬoŋurˈjaeθól
a ship	a sailor	seamanship	nautical

The middle row has the pronunciation in the International Phonetic Alphabet, ˈ marking stress and an acute accent marking rising pitch. It is clear that *-wr* here is a suffix, part of the single word *llongwr*, even though it was in the past an independent word meaning 'man', cognate with Latin *vir*.

Recent research suggests that some of these phonological patterns are important for infants' acquisition of language[10].

Grammatical Criteria

Grammatical criteria can sometimes delimit words: grammatical words. In English nouns typically have a plural form made by adding *-s* to the singular form. We might ask if this *s* is an independent word. The different plural formations, lexically-determined, such as *feet*, *men*, and *children*, are evidence against this hypothesis: if plural *s* was an independent word, we would expect to be able to use it freely to express plurality.

This line of argument is not available with the Japanese case particles discussed above, which are invariant and used with any noun, here exemplified with the object marker を *wo*

子を見る	Someone sees a child
男を見る	Someone sees a man
足を見る	Someone sees a foot
狐を見る	Someone sees a fox
猫を見る	Someone sees a cat
犬を見る	Someone sees a dog

Dixon and Aikhenvald propose that when grammatical elements are necessarily contiguous (nothing can intervene between them), in a fixed order, and with conventionalised coherence and meaning, they form a grammatical word[11]. By these criteria, both English plural *-s* and the Japanese case markers form grammatical words with their nouns.

[8] Nerida Jarkey: 'Words in Japanese', in Aikhenvald, Dixon, & White.

[9] with a tiny minority of exceptions.

[10] "Infants can use rhythmic information ... to guess where one word ends and another begins when listening to natural speech." University of Cambridge Research https://www.cam.ac.uk/research/news/why-reading-nursery-rhymes-and-singing-to-babies-may-help-them-to-learn-language

[11] *Op. cit.*, p. 19.

There is another *s* in English, the genitive *'s*, and here things are less clear. In simple examples genitive *'s* can appear to be a suffix just like plural *-s*, but the following examples are taken from actual text:

the Emperor of Japan's hat
John and Mary's wedding
a guy I know's house
the man she was speaking to's reaction

Here the hat is the Emperor's, the wedding is of both John and Mary, the house is the guy's, and the reaction is the man's. It appears that other words can intervene between genitive *'s* and the possessor to which it relates, suggesting that it is in some sense an independent word. On the other hand it is clearly not a phonological word: it is pronounced as part of the preceding word. This type of element, which grammar suggests is a word but phonology suggests is not, is called a clitic.

Grammatical criteria can also sometimes help to decide whether a word is a lexeme or not. Take the case of the suffix まる *-maru* added to an adjective[12] to make a corresponding inchoative verb in Japanese. On the one hand, metaphorical uses of the adjective may not carry over to inchoatives formed with まる *-maru*. The adjective 高い *taka-i* 'high' can be used to describe a salary, but its inchoative 高まる *taka-maru* 'become higher, heighten, rise' cannot.

給料が高い	The salary is high
給料が高かった	The salary was high
給料が高かろう	The salary may be high
給料が高ければいい	It would be nice if the salary was high
*給料が高まる	*The salary will rise

On the other hand, not all adjectives can take まる. We have 広い 'wide' 広まる 'widen', 弱い 'weak' 弱まる 'weaken', 高い 'high' 高まる 'heighten', 低い 'low' 低まる 'lower', 長い 'long' but ***not*** *長まる 'lengthen'. The lack of consistency in both meaning and applicability suggests that まる-forms are separate lexemes and not wordforms of the corresponding adjective, that they are derived forms rather than inflected forms, to use the traditional terminology.

Intuitions about words

English speakers have strong intuitions about words. Wray[13] quotes Bloomfield:

> *The analysis of linguistic forms into words is familiar to us because we have the custom of leaving spaces between words in our writing and printing. People who have not learned to read and write have some difficulty when, by any chance, they are called upon to make word-divisions.*

— Bloomfield is suggesting that our intuitions derive from our writing system. Dixon and Aikhenvald[14] ask if there are any generalisations that can be made about where people uninfluenced by orthography will place word

[12] I use the English term *adjective* for convenience here; though these words mostly translate as English adjectives, they are grammatically verbs, verbs with present tense in *-i*, called in Japanese 形容動詞 'descriptive verb, adjectival verb'.

[13] Alison Wray: 'Why are we so sure we know what a word is?', pp. 731-2. In *Handbook*, pp. 725-50.

[14] *Op. cit*. p. 30

boundaries. They lament the complete lack of studies of individual languages, but suggest that orthographical word boundaries are in practice inserted at the points where neither phonological nor grammatical words will be split. In many languages, phonological and grammatical words will wholly coincide; in other languages there will be a minority of cases where phonological word and grammatical word do not coincide. Languages which lack phonological or grammatical words, or in which these units are consistently non-coincident, seem unlikely to exist.

Wray asks how Dixon and Aikhenvald's claim might be tested[15]. She quotes conflicting evidence from studies of letters written by semi-literate French- and English-speakers, but suggests that other factors are responsible, and in the end reaches no conclusion on this point. Wray's final conclusion is that language has words but is not exhaustively analysable in terms of words. Concrete nouns are the prototypical examples of words. Less prototypical are 'highly visualisable verbs' and adjectives. 'But at the far end of the continuum there would also be a massive retinue of less well-defined forms, not cut clearly into words'[16].

Japanese is an interesting case in point here. The Japanese writing system does not delimit words, so it might be imagined that Japanese speakers' intuitions about Japanese words could be used to test Bloomfield's and Dixon & Aikhenvald's claims, as well as Wray's prototype idea.

Actually, there are three ways in which the Japanese writing system might prime readers' intuitions about words. Firstly, three types of characters are used: Chinese characters, and two syllabaries. A change from one type of character to another can sometimes seem to be a word boundary; in particular a change from syllabary to Chinese character often coincides with what seems to be the beginning of a word, but this is far from consistent. A second possible source of intuitions about words would be dictionaries. Dictionary headwords, what we called citation forms earlier, are base forms of nouns, present tense forms of verbs (including adjectival verbs), and various particles. Thirdly, picture books for small children are written mostly in syllabary, with few or no Chinese characters, and typically separate groups of syllables with space for ease of reading. The separated groups of syllables tend to be fairly long, phrases after which a reader would naturally pause rather than what an English-speaker might think of as words, but they must surely prime readers' intuitions about suitable points to divide the stream of speech or writing.

I conducted an experiment to try and get at what intuitions Japanese native speakers have about words in their language. All Japanese university students have several years' exposure to English. They know that it is normal when using the Roman alphabet to separate words with spaces. So I simply asked the students of several university classes to transcribe a sentence of Japanese into Roman letters. I gave no explanation of the purpose, saying I would explain later in the class. I told the students to use any Romanisation they liked: it was not a test of the correctness of their Romanisation. A couple of the students used no spaces, several wrote their answer vertically, transcribing one Japanese character per line, but eighty-two gave usable answers, which are analysed below. The students were aged nineteen to twenty-one, sexes in roughly equal numbers. The sentence to be Romanised was projected onto the screen at the front of the classroom in traditional vertical script; the actual prompt is shown at right. (The sentence was the definition of the Japanese *単語* from a standard dictionary, and might be translated as 'individual units of language used as

文を組み立てる要素とし
ての一つひとつの言葉

[15] *Op. cit*. p. 733

[16] *Op. cit*. p. 749.

the basic building blocks of text'.) The results are shown in the table below. The letters used to represent the Japanese sounds have been standardised, so that *youso* includes *yoso*, *yooso*, *yôso*, *yōso*, and *yohso*: it is the use of spaces, not of letters, that we are interested in.

文を	組み立てる	要素としての	一つひとつの	言葉
bunwo^{3} bun wo^{7}	kumitateru^{8} kumitate ru kumi tateru^{1} kumi tate ru	yousotositeno^{2} youso tositeno^{2} yousotosite no yousoto siteno youso to siteno^{2} youso tosite no^{1} youso to site no^{2} you so to si te no	hitotuhitotuno^{2} hitotu hitotuno^{1} hitotuhitotu no^{2} hitotu hitotu no^{4} hito tu hito tu no^{1}	kotoba^{9} koto ba^{1}

There are four places where all respondents put a space, making the five phrases which form the five columns of the table above. Within each column, the superscript numbers show proportions of the respondents spacing as shown, rounded to the nearest tenth. Spacings with no superscript number were used by fewer than five respondents.

It can be seen that there was little consistency. At the top of each of the table's columns is the phrase that all respondents delimited with space. At the bottom of each column are the morphemes making up the phrase, treated as words by at least some of the respondents in each case. In between are various overlapping divisions. There are clearly no universal intuitions about words here, at least if the premiss of this experiment is valid, that the students have in mind the idea that spaces are used to separate words when writing with Roman letters.

The students' responses seem to support Bloomfield's claim that 'the analysis of linguistic forms into words is familiar to us because we have the custom of leaving spaces between words in our writing and printing.' Literate Japanese have no custom of leaving spaces between Japanese words and so no base for dividing Japanese text into words. Wray's idea of nouns and verbs as more prototypical than other parts of speech is not supported: most respondents treated the citation forms 文 *bun* (a noun) and 組み立てる *kumitateru* (a verb) as words, but a large minority did not.

> For those readers not familiar with Japanese, but who are interested in the finer details of this experiment, the following is an explanation of each of the five agreed-upon phrases.
>
> **Bunwo** *bun* is a citation form, a dictionary headword, a noun meaning 'text'. *Wo* is a case particle marking the direct object of a verb (like the Latin accusative case). *Bunwo* is undoubtedly a phonological word, because the accent pattern of a noun includes its case particle, as discussed above. It is probably a grammatical word as well: there is no possibility of inserting a pause or any lexical material between a noun and its case marker.
>
> **Kumitateru** too, is a citation form and dictionary headword. It is a compound verb, meaning 'assemble, construct'. *Kumi* is the infinitive of a verb meaning 'put together'; *tate* is the infinitive of a verb meaning 'build, set up'; both can be used as independent verbs. *Ru* is the present tense morpheme.
>
> **Yousotositeno** *youso* means 'component, element', made up of two morphemes, *you* 'need' and *so* 'ingredient'; *to* is a quotative particle, 'as'; *si* is the infinitive of the verb 'do'; *te* is the continuative

particle, which follows a verbal infinitive; and *no* is the genitive particle, like English apostrophe-s.

Hitotuhitotuno *hito* is the number ‘one’, *tu* is a counter particle. The repetition implies plurality, so ‘one-by-one, singly’. *No* is the genitive particle.

Kotoba is the citation form of a noun meaning ‘speech’ or ‘language’. It is made up of two morphemes, *koto* is language, *ba* is ‘leaf, leaves’.

Conclusion

Nerida Jarkey[17] shows that both phonological and grammatical words are linguistically present in Japanese, and that they mostly coincide. However, there has long been variety in how Japanese text is split into words when this is required. An early example is the sixteenth century publication of Aesop’s fables in Romanised colloquial Japanese[18]. The text is divided into words, separated by spaces, but the words thus delimited are of varying length and type. The Japanese version of Braille, introduced in the mid-nineteenth century, represents Japanese phonetically, each Braille letter being a syllable, and separates words with space. The words thus delimited are basically phonological words, particles included with their nouns, and auxiliaries and verbal particles included with their main verbs. Romanised Japanese dictionaries for English speakers have typically used shorter words, starting with James Hepburn, author of the first published Japanese and English dictionary in 1867[19]. Here and in most subsequent Romanised dictionaries, a Japanese word is basically what would translate as an English word. Plain present and past tense forms of verbs, single words in English, are words in Japanese; but most other nominal and verbal particles are words, including progressive tense forms of verbs, written as two or more words in English, and likewise in Japanese. Computational software for dividing Japanese text into words, such as *Chasen* and *Mecab*, uses dictionary-based words of the same short type. (This software is used in applications such as searching text, machine translation, etc.)

Because of these varying traditions, English-speaking learners of Japanese, linguists, lexicographers, computer scientists, and other specialists, have their own intuitions about what a word is in Japanese, guided by their own specialisms. It may be though that native speakers unschooled in applied linguistics, computer science, or lexicology, have no very clear intuitions about what a word is in their language.

Bibliography

Alexandra Y. Aikhenvald, R. M. W. Dixon, & Nathan M. White (eds.) *Phonological Word and Grammatical Word: A Cross-Linguistic Typology*. Oxford University Press, 2020

R. M. W. Dixon and Alexandra Y. Aikhenvald (eds.) *Word: a cross-linguistic typology*. Cambridge: Cambridge University Press, 2002.

John R. Taylor (ed.) *The Oxford Handbook of the Word*. Oxford : Oxford University Press, 2015.

[17] *Op. cit.*

[18] *Esopono Fabulas*. Amakusa : Society of Jesus, 1593.

[19] *A Japanese and English Dictionary*, by J. C. Hepburn. London : Trübner & Co., 1867.

Abstract

In English and other languages using the Roman alphabet, words are those things which are separated by space in a written text. Do words exist in languages in which they are not visually apparent in text, such as Japanese? Do they exist in unwritten languages? This paper looks at some of the linguistic evidence for the existence of words of various types, in at least some languages. An experiment with Japanese native speakers suggests that they have no intuitive concept of 'word', supporting those who would argue that while the word is a well-founded element of some languages, other languages have no unequivocally corresponding element.

John D. Phillips, Yamaguchi, 15th December 2023