I

Word

A Conceptual Complexity

Introduction

In linguistics, the term 'word' is a conceptual enigma. Although linguistically it is treated as one of the fundamental compositional units of a language, in the history of linguistics, the independent identity of a word has often been questioned, challenged or ignored. Scholars of linguistics have been often reluctant to acknowledge the separate linguistic identity of words because of their varied surface forms and functions within a piece of text.

This leads Matthews (1974) to argue that it is not a word, but a morpheme, which is important in establishing relation of a word with phonology, syntax and semantics. Aronoff (1981) also ignores the separate identity of word as he considers word as nothing but a phonetic string, which is connected to other linguistic entities lying outside the string.

Selkirk (1983), on the other hand, treats word from a purely syntactic point of view as she identifies a set of word formation rules, which are applied to generate words. Bybee (1985) interlinks words with morphology to argue that a word should never be studied free from its meaning, as the meaning of morphemes and contexts determine several linguistic properties used in formal expression of words.

Jensen (1990) argues that it is not words but morphemes, which are primary structural units, and which are typically but not necessarily meaningful. Therefore, it is sensible to focus on morphemes rather than on words in word formation. Spencer (1991) tries to build up an interface that underlies between morphology and phonology to understand linguistic entity of words. His basic

idea of word includes inflectional morphology with an underlying interface between syntax and morphology by which one can explore the processes involved in word formation.

In essence, word is a complex linguistic concept, which is difficult to define in straight terms. Citing different examples from various languages as well as observations of various experts of the field, it can be argued that the concept of *word* is actually interlinked with several linguistic issues such as, pronunciation, lexicology, orthography, morphology, grammar, meaning, derivation, inflection, convention, usage, etc. which directly or indirectly play crucial roles in identification of *words* in a natural language.

Keeping all these issues in mind I have tried to address the problem of word identification. In the following sections I focus on the naïve realization of *word* as found in general conceptualization of common people as well as the concepts presented by traditional, structural, and generative linguistics. After these generalizations, I try to understand 'word' as it is treated as a phonetic unit; an orthographic unit; a morphological unit; a grammatical unit; a semantic unit; as a lexical unit; and as a lexicographic unit.

The naïve realization

From a naïve perspective, a word is a unit of a language that has an orthographic form, has phonetic relevance, carries meaning, and is formed with one or more morphemes, which are linked up, more or less, tightly together. In a typical sense, a word may consist of a root or a stem and may have zero, one or more affixes attached with the root or the stem. Based on the type or process of word formation in a language (inorganic like *Chinese*; incorporating like *Greenlandic* and *Basque*; agglutinating like *Turkish* and *Tamil;* and inflectional like English and Bengali) words may either be kept separate as single lexical units or may be combined together to create larger complex units like idioms, phrases, clauses, and sentences. In naïve realization it is also believed that two or more stems, bases, or words may be joined together to form compound, reduplicated, and portmanteau words.

It is already mentioned that the concept of 'word' is an enigma in the area of linguistics. The assumption that a natural language contains words is taken for granted by most of the people. In general, a word is conceived as a fundamental compositional unit, which in essence, becomes conceptually equivalent to **lexical**

items[1] found to occur in a language. Also the grammar books define other grammatical elements in terms of a word by way of saying that a sentence is a combination of words and that parts-of-speech are classes of words.

However, the problem arises when one tries to define the term 'word' in its strict sense, since based on the surface structure words can sometimes be very difficult to identify or delimit. Although *white space* or *blank space* between two character strings is normally accepted in the writing system of many languages as a distinctive mark of word boundary, languages like Chinese and Japanese do not adhere to this policy. Even in a so-called inflectional language like Bengali, words may contain internal spaces, even if they are nouns, adjectives, adverbs, or proper names, such as the following:

(a) Nouns

ছেলে মেয়ে	(chele meye)	[Noun]	'children'
কাদা জল	(kādā jal)	[Noun]	'muddy water'
দিন রাত	(din rāt)	[Noun]	'day night'
বাস স্ট্যান্ড	(bās sṭyānḍ)	[Noun]	'bus stand'
মরণ দশা	(maraṇ daśā)	[Noun]	'death phase'
কালো মানিক	(kālo mānik)	[Noun]	'black ruby'
লাল বাড়ি	(lāl bāṛi)	[Noun]	'red house', etc.

(b) Adjectives

গঠন মূলক	(gaṭhan mulak)	[Adjective]	'constructive'
নেতা সুলভ	(netā sulabh)	[Adjective]	'leader like'
বৃষ্টি ধৌত	(bṛṣṭi dhauta)	[Adjective]	'rain washed'
বালুকা বিস্তীর্ণ	(bālukā bistīrṇa)	[Adjective]	'spread with sand'
ভক্তি বিহ্বল	(bhakti bihval)	[Adjective]	'emotional'
লাভ সংক্রান্ত	(lābh saṃkrānta)	[Adjective]	'relating to profit'
যুক্তি সঙ্গত	(ẏukti saṅgata)	[Adjective]	'logical'
অবস্থা সম্পন্ন	(abasthā sampanna)	[Adjective]	'rich with money'

[1] Lexical items are single word or (grouped words) in the lexicon of a natural language. For instance, *cat, traffic light, take care of, by the way,* and *don't count your chickens before they are hatched*, etc. are considered as single lexical items. Lexical items are generally understood to convey a single meaning, much as a lexeme, but are not limited to single word units. Lexical items are like *semes* in the sense that they are *natural units* translating between languages, or in learning a new language. In this last sense, it is sometimes said that language consists of grammaticalized lexis, and not lexicalized grammar. The entire store of lexical items in a natural language is called its lexis or lexicon.

(c) Adverbs

কোন ক্রমে	(kona krame)	[Adverb]	'any how'
নিদেন পক্ষে	(niden pakṣe)	[Adverb]	'minimally'
বিশেষ ভাবে	(biśeṣ bhābe)	[Adverb]	'specially'
ভুল বশত	(bhul baśata)	[Adverb]	'by mistake', etc.

(d) Proper names

তিতাস একটি নদীর নাম	(titās ekṭi nadīr nām)	'Titas is the name of a River'
যারা বৃষ্টিতে ভিজেছিল	(ýārā bṛṣṭite bhijechila)	'Those who got drenched in rain'
বিকেলে ভোরের ফুল	(bikele bhorer phul)	'Morning flower in the afternoon'
যখন সন্ধ্যা নামল	(ýakhan sandhyā nāmla)	'When the evening set in'
তিস্তা পারের বৃত্তান্ত	(tistā pārer bṛttānta)	'History of the bank of Tista'
দিকশূন্যপুরের পাখী	(dikśūnyapurer pākhī)	'Bird of Dikshunyapur'
বিষন্ন সন্ধ্যার আলো	(biṣaṇṇa sandhyār ālo)	'Light of a sad evening', etc.

When one looks at the example given in (d), one is invariably confused to decide whether these are words, phrases or sentences. These are mainly the names of some literary works (mostly Bengali fictions), which are actually made of several words. However, at the time of linguistic analysis (e.g., at the time of adding case markers or inflection or enclitics, etc.), the entire multi-word string or phrase is treated as a single word unit. Therefore, the confusion is still there in identification of these proper names – should these be treated as single word units or multi-word units.

A synthetic language like *English*, on the other hand, combines together several different pieces of lexical forms into single words, making it difficult to separate them in traditional sense of words found in the analytic languages. In a synthetic language, a single word stem (e.g., *love*) may have several different forms (e.g., *love, loves, loving*, *lovable, lovely, lover, loved*, and *beloved,* etc.). Normally, these forms are not considered as different words, but different forms of the same word. Thus, in languages like English, wordforms are normally constructed by application of a permitted sequential combination of a number of candidate morphemes, which act as word – formative components (e.g., *love, -s, be-, -able, -er, -ed, -ly,* etc.), as the following examples show:

(a)	Word	: love	
(b)	Word-form	: loves	[< love + -s]
(c)	Word-form	: loving	[< love + -ing]
(d)	Word-form	: lovable	[< love + -able]
(e)	Word-form	: lovely	[< love + -ly]
(f)	Word-form	: lover	[< love + -er]

(g) Word-form : loved [< love + -ed]
(h) Word-form : beloved [< -be + love + -ed]

The process of identification of words becomes far more problematic for the polysynthetic languages, such as, Inuktitut[2], Ubykh[3], Eskimo, etc., where a single word string stands for a sentence. In these languages, where space does not necessarily indicate word boundary, word boundaries are normally determined by close reference to the context of a piece of a text. For instance, in *Greenlandic* language the form *aulisariartorasuarpok* means 'he hastens to go fishing', where the *sentence-word* is made of the following components tagged together:

(a) *aulisar* 'to fish'
(b) *peartor* 'to be engaged in'
(c) *pinnesuarpok* 'he hastens'

These polysynthetic languages are often known as **incorporating languages,** where we get amalgamation of the most significant sounds of those different sense elements, which would, in most other languages, stand as separate words. Therefore, it is fair to assume that the main characteristic features of the polysynthetic languages include 'sentence-word' and 'dropping of one or more syllables of each component when these formative elements are incorporated'. It often happens that individual components have merely got a hypothetical existence, and are never actually used alone as individual words in the language, except in some of the more advanced types. The head name **holophrastic** is also given to this type of languages because the whole situation is expressed literally through one word or a phrase. In essence, polysynthetic languages incorporate everything (i.e., subject, verb, object, and all adjuncts) into one word.

There are other languages, which deploy independent words and follow the strategies for sentence construction according to their own methods. However, at the same time, these languages incorporate, although in certain cases only, the pronominal elements also. Thus, in *Basque,* we have the pronoun incorporation both for the subject as well as for the object. In this language the verb proper

[2] Inuktitut (or Eastern Canadian Inuktitut) is the general name of some of the Inuit languages spoken in Canada. It is spoken in all areas north of the tree line, including parts of the provinces of Newfoundland and Labrador, Quebec, to some extent in North Eastern Manitoba as well as the territories of Nunavut, Northwest Territories, and traditionally on the Arctic Ocean coast of Yukon. It is recognized as an official language in Nunavut and the Northwest Territories.

[3] Ubykh or Ubyx is an extinct Northwest Caucasian language once spoken by the Ubykh people who originally lived along the eastern coast of the Black Sea before migrating to Turkey in the 1860s.

has no existence apart from its pronominal complements as the following examples show:

(a)	da-*kar*-kiot	'I carry it to him'
(b)	na-*kar*-su	'you carry me'
(c)	ha-*kar*-t	'I carry you', etc.

Word in traditional linguistics

In ancient Greece, Aristotle gave a formal definition of the word as linguistic unit, which is primarily a component of the sentence, and which has a meaning of its own and is not further divisible into meaningful units (Robbins, 1967). However, according to scholars, Aristotle's definition of word is not adequate, since it excludes the morpheme from consideration, which is always capable of grammatical function, as it carries an isolable meaning (Robbins, 1967). Plato, on the other hand, had not explicated whether his concepts like *onoma* (literally means 'name'), and *rhema* (literally means an 'utterance' or 'thing said') referred to words or to phrases or to both the elements.

In early Chinese linguistic discussions, some discussions were made to make distinctions between the '**full words**', which were capable of standing alone and bearing a individual lexical gloss and the '**empty words**' (i.e., particles), which primarily served different grammatical functions within sentences containing full words. These empty words' scarcely had a separate meaning in isolation, as these were mainly grammatical elements. The full words were further divided into 'living words' (i.e., verbs) and 'dead words' (i.e. nouns) (Robbins, 1967).

In ancient India, logicians and grammarians debated on the question of primacy of the word as against that of the sentence. They argued the extent to which meanings could be regarded as a natural property of words so that one-to-one relationship may be established between a word and the meaning it denotes. Also, they debated for long to understand if words primarily denoted particulars, classes, or abstract universals. Furthermore, there were debates to justify how far word meanings were positive in identifying an object for what it was or negative in distinguishing it from the rest of the reality (Robbins, 1967).

The classical phoneme theory accepted word boundaries as legitimate properties for identification of words in speech. It focused on consonant and vowel segments and tones (in case of tonal languages) to identify words (Robbins, 1967).

The strategies adopted by twentieth century linguists to deal with the problem of word are also diversified. For them words are unique linguistic entities, which bear no phonetic – semantic resemblance to any other linguistic forms available within a language (Bloomfield, 1933). These are meaningful linguistic forms, which are usually accessed and analysed for understanding a natural language. Since these are meaningful linguistic units, these may be accessed to investigate the difficulties involved in lexical productivity, where morphemes will play a crucial role in the act of word formation (Hockett, 1958). In essence, words do not differ much from sentences, since these do not differ fundamentally from any other syntactic units. For instance, it is possible to build up an interface between morphology and phonology to interpret words if it is assumed that a word actually relates to phonology with an interface lying between syntax and morphology (Spencer, 1991). On the other hand, it is also possible to emphasize on fundamental morphological notions to understand a word, as a word still remains a *non-entity*, which is yet to come out from the sphere of phonology, morphology, semantics, and syntax to establish its independent linguistic identity. Thus, many linguists are reluctant to admit independent entity of words and are actually willing to interpret it with phonology, morphology, syntax, or semantics (Katamba, 1993).

The analysis of orthographic forms of words has been a debatable issue in linguistics for generations. A word, due to its unique identity, may refer to a string of characters (or letters) as it appears in writing. Also it may refer to a more abstract entity – as a linguistic unit as it is observed in lexicographic works like dictionaries, and thesauruses. In this context, words are different from sentences because structures of words are much varied than that of sentences. Besides, there are some principles that govern the structure of complex words and these principles are normally applied to form different words of different lexical classes (Aronoff, 1981). Moreover, once words are formed, these become open for orthographic-cum-semantic alternations over time in a language. Gradually, they take on to idiosyncrasies with the result that they become no longer possible to generate by way of simple algorithm of word generation. As word formation process is based on words, the application of word formation rules on existing words is capable of generating new words. Thus, both new and existing words become members of the major lexical categories of a natural language (Aronoff, 1981).

There are, however, some specific and unique properties that help in distinguishing words from morphemes and phrases. In a wider sense, since words are *referentially opaque*, it is, therefore, impossible to *see inside* them and

refer to their compositional parts (Spencer, 1991). On the other hand, from the syntactic point of view, the rules of syntax tend to consider words as the smallest meaningful linguistic units, which may be combined together for composing larger constructions like phrases and sentences. In this frame, words are nothing more than those *minimal free forms,* which exist on their own irrespective to the components (i.e., morphemes) used for their composition.

Within the traditional model, it is understood that the meaning of a word is not always possible to determine compositionally. At certain times, a word may carry an apparent meaning, while in other cases, the relationship between the meanings of the parts and the meaning of the whole word is quite obscured. Also, there is considerable difficulty in finding out a universally applicable notion of word with respect to its form and meaning, even when form and meaning of the constituting parts are taken into consideration (Spencer, 1991). Therefore, understanding a word in terms of various grammatical criteria (e.g., *roots, stems, marker, suffix,* etc.) is a difficult task, since these criteria may sometimes become deceptive in form and function even within a single language.

It is observed that each natural language does possess a well-defined frame of grammar for its word structure which, nonetheless, conforms to certain general principles that govern the possible structure of words in the language (Selkirk, 1983). Once words are put in the lexicon, the morphemes out of which words are formed and into which these are to be analysed, do not have any constant meaning, and in some cases, have no meaning at all. Here begins the problem in conceptualizing words, since words, even if these are formed by some regular word formation rules, can change morpho-semantically. Therefore, it becomes difficult to categorize the meaning of individual words in a principled manner.

To overcome this problem it has been argued that it is always better to have a dictionary, which should contain actual words as well as all their idiosyncratic variants, since words are able to mean more than one thing in different contexts and situations of their occurrences (Halle and Chomsky, 1968). The list of idiosyncrasies should include all phonological and syntactic *exception features,* which are not provided and defined by simple general rules of word formation in morphology. This argument, however, has some limitations, since it fails to show which words are so idiosyncratic that their meanings are totally divorced from what these are actually expected to mean by simple general morphological rules. Therefore, it is difficult to find out how these can mean something different from their expected meanings without damaging their rules of generation.

Word as a phonetic/phonological unit

The most confusing part in identifying words lies in spoken form of a natural language. A spoken text provides only a few phono-lexical cues to identify where the actual boundaries of words exist. Therefore, identifying individual words in a normal speech sequence is a real complex task, since while short words sometimes run together, long words are often broken up into smaller units. Since most of the languages are endowed with this feature, systematic determination of word boundaries in spoken texts is a real challenge. However, the five following processes can be applied to determine where the word boundaries of a spoken text should be placed in:

(a) **Potential pause:** Here a speaker can be asked to repeat a sentence slowly, allowing for pauses so that the speaker is allowed to insert pauses at word boundaries. However, this method is not a foolproof one, since a speaker can easily break up the polysyllabic words into several small parts to make it a string of several words.

(b) **Indivisibility:** Here a speaker is requested to pronounce a sentence loud. After this, he is requested again to say the same sentence again with extra words added to it. Thus, *I am living in this village for last ten years* might become *I and my family are living in this little village for about ten or more years*. Additional words are normally placed at word boundaries of the original sentence. This process is not optimally effective, since some languages have infixes, which are embedded inside words while other languages have separable affixes detached from words. For instance, in *Hindi* a sentence like *tum ýā rahe the*, the verb is split into three components (i.e., *ýā, rahe* and *the*) because it is a continuous form of the verb which is not possible to express as a single word-unit in Hindi. Therefore, these forms are used sequentially as separate units in the sentence. Similarly, in *Bengali*, the noun ভালোবাসা (bhālobāsā), in some situations, is split into two parts [i.e., ভালো (bhālo) and বাসা (bāsā)] and are placed at two different places within a sentence, as in, ভালো যে তুমি আমায় কতই বাসো, তা আমিই জানি (bhālo ýe tumi āmāy katai bāso, tā āmii jāni) "I know how much you love me!" This implies that indivisibility cannot always be a dependable criterion for identification of words in a language.

(c) **Minimal free forms:** This approach was first reported to be adopted by Leonard Bloomfield on the basis of the concept that words are the smallest meaningful units of speech and that they can stand by themselves in a language. This concept on the one hand actually correlates to phonemes, which are treated as the distinct units of sound, and, on the other hand,

correlates to lexemes that are considered as the primary units of meaning. This approach, however, fails to solve the problem of identifying words as 'minimal free forms', since in many languages many written words are not at all minimal free forms as they fail to make any sense by themselves. For instance, English forms like *the, a, of, by, in, up,* etc. are not free forms in the true sense because if isolated from context of their usages, they fail to denote any sense or meaning.

(d) **Phonetic boundaries**: Some languages have specific rules of pronunciation that may help identify boundary of words. For instance, in a language, where the word-final syllable is regularly stressed, a word boundary is likely to occur after each stressed syllable. Another example can be observed within a language that has vowel harmony (e.g., Turkish) where the vowels within a given word share the same *quality*. In that case, a word boundary is likely to occur whenever the quality of the vowel is found to be changed. Since all the natural languages do not have such convenient phonetic rules or conventions, this feature may be treated as occasional exceptions, applicable to a few languages that adorn this feature.

(e) **Semantic units:** Here the basic argument is that a word is a linguistic unit, which has its own phonological and orthographic (in case of languages having script or writing system) form with a separate semantic identity. Similar to the argument of *minimal free form*, this method also fails to break down sentences into some smallest units like words having separate semantic identity. It is noted that most of the languages often contain many words that have little semantic specification, as they often play specific grammatical roles than playing any specific semantic role. Moreover, it is noted that in the case of compound words (particularly in the case of exocentric compounds) the semantic specification of individual words is often lost to generate different meanings.

To overcome this problem, linguists often tend to combine all the strategies to determine word boundaries of any given sentence. Even with careful application of these methods, the exact definition of a word may remain elusive, since there are words (e.g., *head, right, soon, donkey, hero, flight,* etc.) that appear denotative but are actually connotative due to several factors controlling their occurrences in several contexts of a language.

Word as an orthographic unit

From an orthographic point of view, a word is treated as a sequence of letters or characters bound together within a single string, which has a white space at each