On Defining Word

Kavi Narayana Murthy

School of Computer and Information Sciences

and

Department of Sanskrit Studies

University of Hyderabad

email: knmuh@yahoo.com

Abstract

In this paper we shall explore one of the most fundamental questions in linguistics, namely, what exactly is a word. In Natural Language Processing and related technological areas, a word is invariably taken as a sequence of written letters separated by spaces. Linguists do understand the distinction between language and script and the natural predominance of the spoken form over the written form. Nonetheless, it appears that the written form has had a profound influence on the way linguists actually work, especially in modern times. In this paper we highlight the problems and issues with such a perspective and we argue for a more semantically motivated and hence more universal definition of a word. Our goal here is not to give one final answer to the question of what constitutes a word but to argue for a better choice than what seems to be the accepted working definition today. The foundations we shall give here are a result of many years of deep study in our group here at the School of Computer and Information Sciences, University of Hyderabad, India. Since the paper touches upon a number of fundamental questions of great relevance to various branches of linguistics including applied areas such as language technologies, the paper has been written in a simple, clear, tutorial-like style.

1. Introduction

Before we get to the main question of what counts as a word, it is necessary to build up the required background. Computational Linguistics, also known as Natural Language

Processing (NLP), started off in the mid fifties, as a part of Artificial Intelligence (AI), at about the same time when modern generative linguistics was taking shape. AI is a discipline with two broad goals - 1) to understand the nature of human intelligence and 2) to build intelligent systems. Language is at the very core of human intelligence and understanding the nature of human language faculty thus becomes the most important goal of NLP. Once sound foundations are laid, technology and applications naturally follow. The primary goal of NLP is to understand how exactly we human beings understand, produce and learn languages. In NLP, also variously known as computational linguistics, language technology and human language computing, we use computing machines to model and experiment with the human language faculty. Otherwise, it is not really different from linguistics per se. The core ideas we present below are not all entirely new, but their precise formulation and presentation is our own.

1.1 Language

Computer science defines a *formal language* as a set of strings defined over a finite set of discrete atomic symbols. While this definition is precise and suitable for certain kinds of analyses, it does not directly reflect the nature of human languages or the nature of the human mind. Therefore, this definition does not suit our purpose here. We should also be careful not to confuse particular languages such as English with language per se. Here language is to be understood as a faculty of the human mind.

Let us look at the mental capabilities of human beings related to language. Firstly, we humans are capable of producing a number of different kinds of sounds under our full control and will. Secondly, We are capable of making interesting patterns by stringing together these basic sound units, (called *varNa-s* or *phonemes*), into larger and more complex structures such as words and sentences. Thirdly, and, most importantly, we are capable of systematically associating meanings with these patterns of sounds. Fourthly, we are capable of communicating our ideas, thoughts, feelings and emotions to others by expressing them in patterns of sounds according to these structure-to-meaning mapping rules. Fifthly, we are capable of using the same capabilities for deliberative thinking (See the book entitled "Freedom" by the same author (Murthy, 2012) to understand this better.) Sixthly, we are capable of learning such sound-meaning associations and we are capable of teaching these association rules to others too. We shall call this six-fold faculty of the human mind as language. *Language is the capacity of the human mind to systematically map sounds to meanings and use this for speech and thought*. *Meaning is the core idea.*

Language is a faculty of the human mind. As such, it is not an object, it is not a thing. Language is not a set of valid words or valid sentences. Nor have we defined any particular language here. Also, we have chosen to exclude animal language, machine language, gestures and body language, sign language etc. from our definition intentionally. Speech is the most basic form and other varieties of language can at best be substitutes or alternatives, suitable or necessary only in certain special circumstances. Language is our mental faculty, we have invented sign language as an alternative when speaking and listening are difficult. Sign language is not a natural mental ability, it is invented by us, it needs to be taught and learned. Similarly, machine languages have nothing to do with human mental faculties. By the same logic, writing is not a natural mental ability and we need to keep it aside. Therefore, language and every aspect of it should be viewed and understood from the point of speech, not writing.

1.2 Grammar

One of the core challenges in the scientific treatment of language is discovering the mapping between form and function (or structure and meaning), and reconciling anomalies if any. Thus we can consider language to be a system for mapping sound patterns to meanings. How do we do this mapping? Unless we have a very systematic and principled way of mapping sounds to meanings, we will not be able to teach, learn, or use language for thinking or communication in an orderly, efficient, predictable and purposeful way. *Systematization of the body of rules and principles for mapping structure to meaning is what we shall call grammar. The fact that we are all able to use our language in our daily life proves that we all have a grammar of our language in our heads. Grammar is not what you find in some grammar book, the mental grammar is the real grammar and discovering this grammar is the main task of linguists*. The goal of any formal grammar should be to model the human mental grammar as closely as possible.

Grammar is the essence of linguistics. Grammars that map structure to meaning at the level of words, sentences and discourse, form the central core of modern linguistics. *We need to define a word, we need to understand the structure of words and the relationship between this word internal structure and word meaning. This component of grammar is called morphology. Then we need to define a sentence. We need to define the relations between the* *words in a sentence. This way we can understand the structure of sentences and the relationship between sentence structure and sentence meaning. This we shall call call syntax*. Morphology and syntax form two of the most important components of a computational grammar.

Words may include two parts. One part, which we may call the root or base, gives the lexical meaning, while the other part, which may be expressed through affixes, gives the grammatical part of meaning. Thus in the word 'trees', 'tree' is the base form and '-s' is an added suffix indicating the plural value for the grammatical feature called number. Morphology tells us how to compute the meaning of the word 'trees' provided we know the meaning of 'tree'. The meaning of the base 'tree' cannot be computed from its parts, that is, from the phonemes or characters it is made up of. In general, the mapping between roots/bases and meanings is arbitrary and the only way to specify this relation is by explicit listing. The human mind stores words. To be more precise, it stores the mappings from sound patterns to meanings. This is called the mental lexicon. We have defined grammar as a way of mapping form to function. Therefore, the lexicon is also very much a part of grammar, not an adjunct to it.

There can be more than one way to understand the structure of a given linguistic unit and how it can be mapped to meanings. Therefore, there can be many valid grammars that correctly describe a given language. The goal is to discover the simplest, the most natural, the most elegant, and the most efficient grammar.

A baby born in any language community anywhere in the world picks up its mother tongue in more or less the same amount of time and with equal ease. Therefore, although various human languages appear to be quite different from each other, they must all be essentially the same at some sufficiently abstract level. Also, we see no major differences across the peoples of the world in other mental faculties such as perceiving, reasoning, planning or decision making. The human mind is the same everywhere. Why should languages be different? There must be some universal principles underlying all human languages. The primary goal of modern linguistics is to unearth these universal principles underlying all human languages so that a Universal Grammar can be developed. Individual languages such as English, Hindi and Telugu are specific instances of such a universal grammar.

Linguists generally develop the subject by considering how valid linguistic expressions such as words and sentences can be constructed or generated from a given abstract grammar. Text book grammars also generally take the same view point. In NLP the general tendency is to look at how already constructed words or sentences can be analyzed and understood. *Whether we take the generation point of view or analysis point of view is a matter of choice but it is important to realize that there is only one grammar in our mind and the same grammar is used both for analysis and generation*. There is no clear evidence to show that the human mind uses two different grammars, one for generating linguistic utterances and the other for analyzing and understanding what others say. That would be very unintelligent and wasteful. Even in computers, we must therefore look for grammars that can be easily and efficiently used for both analysis and generation.

1.3 The Language Code

When a speaker and a listener communicate through speech, what goes from the speaker to the listener? Words do not go, sentences do not go, meanings do not go, even the air does not go. At each point in the medium of air that separates the two persons, air pressure varies with time but actually no material stuff travels or goes from one to the other. A small part of the physical energy of the spoken sounds does reach the ear of the listener but energy itself has nothing to do with language or meanings. Yet this somehow facilitates communication of meanings. *Language is a unique kind of bridge, which connects two minds and enables communication of meanings but the meanings themselves never cross this bridge. Meanings reside within the human minds, before, during and after the utterance*. This is a critical point to be noted. To understand this section fully, interested readers are advised to go through the book entitled "Freedom" by the same author (Murthy, 2012).

What goes over the bridge are mere triggers. The speaker maps meanings to symbols that have the potential to trigger the same meanings in the mind of the listener. These triggers travel through the medium in the form of sound energy and once the listener gets these symbolic triggers, these triggers invoke the corresponding meanings inside his own mind. *There is no meaning in text directly, there is no meaning in speech directly, there is no meaning in the air, yet language facilitates communication of meanings, feelings, emotions, ideas and thoughts between human beings. This is the beauty of language*.

The meaning structures inside our mind can be non-linear and quite complex but linguistic expressions are always simple, linear sequences of symbols. Speech is a sequence of basic sound units, called phonemes, arranged linearly along the time line. Phonemes are uttered one after the other, they are heard one after the other in a temporal sequence. *Language is a scheme for coding complex non-linear, non-symbolic meaning structures into a linear* *sequence of symbols and for decoding this back into the original meaning structures. Each human language is a different system for coding and decoding the same kinds of mental structures*. (Of course, the written text is also a linear sequence of symbols but as we have already agreed, we shall not worry about the written form here.) *Language works by encoding and decoding of meanings indirectly through symbolic triggers*. Understanding the principles of such encoding and decoding is the central task of linguistics and NLP.

Because there can be more than one way of relating sound patterns to meaning structures, there can be more than one language. Fine, but this still does not say why there are actually so many different languages the world. Phoneme sets used by various human languages are largely common. Words show maximum variability. Significant differences are seen at the sentence level too but when viewed from the right perspective, we will begin to realize that syntax is actually quite uniform across languages. As we move towards the discourse level, everything becomes a lot more common and universal. After all, it is the same human mind which is the king pin in this whole business of language. Languages are not as diverse as they appear to be.

Why do we see so much of difference between languages at the level of words? Why do languages differ so much in terms of morphology? Actually, the differences are not so much. If we take the right approach to words, most of these differences will melt away. This is the central thesis of this paper.

1.4 Computing Language

Our focus here is on language and computation. Computers are purely symbolic machines, devoid of all feelings and emotions. A computer can store and process symbolic structures standing for words, sentences and bigger discourse structures, but it can never really get any meanings. Computers will never be able to understand the meaning of Natural Language utterances in the sense we humans can, nor will they ever be able to communicate any feelings, emotions or meanings with human beings or other computers. They can simulate human behaviour, but they cannot actually experience. How then do we do any meaningful work on language using computers?

The answer to this critical question comes by observing that there is structure in language and structure is systematically related to meanings. Computers can represent linguistic units symbolically and manipulate these stored structures in interesting and useful ways. If only we take care to see that these stored structures remain faithful to meaning *at* *all stages,* we can get a great deal of interesting, useful and meaningful work done using computers.

2. Issues with the Prevailing Views and Practices

Since the beginnings of grammatical study in Europe, the concept of a word has been considered to be of central importance. There have been several attempts to define a word. Attempts have been made to define a word starting from concepts of orthography, phonology, lexeme, morpheme and syllable. All these definitions have one major limitation. They do not necessarily lead to meaning in a natural, straight-forward and language independent way. Our goal is to develop a universal computational system to compute the meaning of words and sentences. Further, there is a large gap between theory and practice. Whatever we may say about words in theory, when it comes to practice, we find that tokenization on spaces is almost always resorted to.

In India, we have a very rich grammatical tradition going back to several thousand years. In particular, three major schools are relevant for study of language - nyaaya (logic), miimaamsaa (jurisprudence) and vyaakaraNa (grammar). Our work here is influenced by our studies in all these areas. Speech is given primary importance, not writing, in Indian tradition.

Consider the following English sentence and its rough equivalents in Telugu, Hindi and Kannada. We use double letters for long vowels, upper case letters to indicate retroflex sounds and an added 'h' to indicate aspiration. The upper case letter M stands for the anusvaara in the written form. These conventions are quite standard and well known. See (Murthy and Srinivasu, 2012b) for more on the Romanization scheme used in this paper.

1. The dog had been running for some time (English)
2. kukka koMta seepu parugeDutuu uMdi (Telugu)
3. kutta thooDi deer kee liyee dauD rahaa thaa (Hindi)
4. naayi kelakaala ooDuttittu (Kannada)

How many words are there in these sentences? Both in NLP and in modern linguistics, the usual answer would be 8, 5, 8 and 3 words respectively. The four sentences are rough translations of each other and the overall syntactic structure as also the meaning is the same in all. We are talking about one entity – a dog. We are talking about one kriyaa (action) performed by the dog. We are saying something about the time of this kriyaa. We need one noun to indicate the dog, one verb to indicate the action and one adverb of time to indicate time. Why then do we have different number of words in these sentences? What is the extra thing we have in English or Hindi compared to Telugu or Kannada? Is something missing in the Kannada sentence? What is really common to all these four languages and how do we explain the differences despite this common core? These are the main concerns for us in this paper.

In computer science, terms such as characters, character strings or words, sentences, grammar, and language are defined keeping the written form in mind. NLP normally deals only with written texts and naturally these same definitions are carried through into NLP. This is not right. Seven and a half Billion people across the globe carry on their daily life by transacting in speech. People speak words and sentences and they hear words and sentences. There are no characters or spaces in speech. How can we define words as sequences of characters delimited by spaces? Are there no words in speech?

Definitions that take the written form as the basis suffer from many defects. Firstly, there are many languages in the world which do not have a script at all even today. These languages are spoken, heard, understood, used effectively like any other language. Ordinary people can use these languages for speech and thought as effectively as any other language. Therefore, writing is not an essential aspect of language. People were speaking, listening, thinking and using language for thousands of years before writing was invented. Secondly, definitions based on the written form cannot be applied directly to spoken form of language, whether the language in question can be written down or not. Thirdly, we learn the spoken language naturally, automatically, without being taught, right in our early days. We learn reading and writing much later, perhaps in school, by being taught. Reading and writing are technologies, they are inventions of mankind, they are not natural abilities, they depend upon external material resources such as books and pens. Speaking and listening are the most natural, effortless, efficient activities, leaving the hands and eyes free to do other things. Reading and writing are not natural and innate to our mind. Our focus here is on the innate capabilities of the human mind. It would be very bad, therefore, to define terms like language, grammar, word and sentence based on the written form. We should not let the written form of language influence our thinking. Fourthly, even in languages which have a script, many people remain illiterates. Illiterates need not be ignorant or foolish, in fact they can be great scholars. In India, writing has always been considered as an impediment and looked down upon. Only those who are not smart enough to remember things write down. In olden times, the greatest of the scholars and scientists of the day chose to remain like illiterates voluntarily, staying away from writing down anything. Intelligence comes from careful listening. Reading and writing makes one dull and mechanical. Reading and writing in fact only tax the mind and interfere with its natural working. Reading and writing are also wasteful of material resources such as paper and ink. It is possible to remember entire books by heart and Indians have mastered the art of preparing texts which can be easily memorized and remembered for life. There are a number of techniques to help people use language only orally, and still ensure that not a single bit of information is lost, distorted or confused. Literacy is quite a hollow concept, not as important as people think it is. This has been the traditional Indian view. Fifthly, even after the invention of writing systems, most languages were written without any space between words. Stone inscriptions etc. show no spaces between words. You may not find spaces between every two words in poetry. In many languages/scripts, words are written without spaces even today. Writing spaces between every two words is a very recent phenomenon and no one has bothered to define precisely where one should insert a space and where one should not. The written form is quite arbitrary, ad-hoc and lacks any sound basis in most languages of the world. Writing conventions vary from language and language and there is often a good degree of inconsistency even within a given language. Sixthly, a language can be written in any script and one script can be used to write any language. This is not just a hypothetical possibility, many languages and scripts are in fact written in various combinations this way. Which script should we take as the basis for defining words? Finally, meaning is supreme in language, if we sideline meaning, the entire exercise becomes meaningless. In most scripts in the world (except in principle, in the so called ideographic scripts, where the intention is to render units of meanings directly), the relation between the written form and meaning is not direct and one-to-one. How can we go by the written form? Given all this, defining words as sequences of characters (that is, letters, symbols, etc.) separated by spaces is not at all a good choice we can make. Let us not reduce linguistics and NLP to arbitrary manipulation of written symbols. Language, word, sentence, structure, grammar, meaning, all of these can be and should be defined and dealt with, without getting influenced in the least by the written form.

NLP deals with written texts and the standard practice has been to divide a given piece of text into units based on intervening spaces and take these units as words. The words we get this way do not always correspond to units of meaning and are thus not universal. Compounds, external sandhi, the so called multi-word-expressions, etc. lead to avoidable confusions. The grammatical categories and properties we need to introduce will also not be universal and semantically motivated. Prepositions, for example, are not universal, nor do they correspond to units of meaning in a simple and direct manner. What should be considered as one word may get split into many, causing a variety of problems. For example, treating 'has been running' as three words instead of one, leads to strange notions such as auxiliary verbs. If verbs are words that indicate action, and there is a single action here, how can there be three words (categorized as verbs) here? Do auxiliary verbs indicate auxiliary actions? Also, aspects of a verb, such as tense, aspect and agreement get split across units in complex and possibly even inconsistent ways. It becomes difficult to develop simple, elegant, natural, efficient grammars. A big part of the load of morphology gets into syntax, making syntax too complex and unwieldy. Structural ambiguities multiply. Words and sentences appear to be highly ambiguous and the central theme of NLP is taken to be dealing with these ambiguities. Underlying universals are blurred and languages appear to be greatly different from one another. Grammars become large, complex and unwieldy. Meaning takes a back seat and the whole field degenerates into mere string manipulation. New layers of processing are invented to deal with the high degree of ambiguity. POS tagging, chunking, shallow or partial parsing are examples of this. There does not appear to be any strong theoretical motivation to introduce such layers either from the point of view of linguistics or from the point of view of AI. New terminology is created, for example – multi-word expression and local word grouping, which do not seem to have any strong scientific foundations. NLP becomes more of art and less of science. We believe that the root cause of all these tendencies is the lack of clarity on what is a word. The central goal of this paper is to propose an alternative and better view of the concept of a word.

Only valid words must enter the lexicon. Only valid word forms must be sent to the morphology for analysis. Tags should be assigned only to words, not to tokens of orthography. Sentences should be analyzed in terms of words, not orthographic tokens. Then our understanding of related concepts such as morphemes, bound and free morphemes, word classes and grammatical categories, etc. will also change.

3. The Concept of a Word:

When we think of some word, say, 'tree', in English, there are several things that are relevant. We could think of an individual physical object existing in the external material world or a collection of such objects denoted by the word tree. Physical existence in the outside world is not a necessity and so we can think and talk about concepts such as anger and love as also about imaginary things such as a flying elephant. Here we shall not bother about external worldly reality at all. Secondly, we have a meaning for this word tree. Then we have the sound, that is, the phoneme sequence that will be heard when this particular word is spoken. Then, we may also have a written form, the letters 't', 'r', 'e' and 'e' written one next to the other without any intervening space.

Pronunciation and meaning are the two essential and most important properties of a word. The relation between these two is the crux of the matter. How the word is written – its spelling, is secondary. We should not be thinking in terms of letters, alphabets, characters, symbols, scripts, fonts, glyphs etc. We should think only in terms of phonemes, phoneme sequences and meaning.

A good way to understand any entity is to check what it is made up of and what it is a part of. Words map sound patterns to meanings. In the spoken form, a word is represented as a sequence of basic sound units called phonemes. We need to understand how words are built from phonemes. Words join hands to build sentences. We need to understand how words relate to one another to give meaning to a sentence. Working from both ends, we can fix a word precisely.

3.1 Phonemes: The Building Blocks

Phonemes are the basic sound units in a language. They are called varNa-s in our tradition. /a/, /b/, /t/, /k/, /m/ are examples of phonemes. *Phonemes are abstractions, not exact sounds*. Various phonetic realizations, or phones and allophones may be possible but only those differences which affect the meanings of words are important in phonemics.

Each language has its own set of phonemes, and two sounds, which form different phonemes in one language can actually be a single phoneme in another language. Thus /g/ and /k/ are different phonemes in Kannada but they merge into a single phoneme in Tamil. *Human languages have somewhere between 30 and 50 phonemes and the phoneme sets of different human languages share a large common intersection*. Languages do not vary vastly in terms of the phonemes they use, the differences are small, the uniformity is glaring.

To characterize any language, the very first step is to list the phonemes used in that language. The phonetic status of a candidate sound is established through minimal pairs. It is important to note that *the phonemic status of a candidate phoneme can only be established in the context of the whole phoneme system, not in isolation*.

3.2 From Phonemes to Words

Phonemes are small in number, less than fifty or so. Since we are capable of expressing a much larger and richer set of meanings, languages normally do not map phonemes to meanings directly. *The smallest units of meaningful expression are not individual phonemes but short sequences of phonemes, called words. In this sense, words are the minimal meaningful units in a language*. Words can be used to build larger units of meaning such as sentences and discourse segments.

How many words are there in a given language? Often we have no clear idea. This happens not only because we may not have explored this question at all, but also because we do not have a clear idea of what exactly constitutes a word. The number of words in a given language may be in tens of thousands or lakhs or even crores. Yet *the set of words in a given language is finite*, never infinite. New words may come into a language, new words may be constructed through productive word formation processes, but still it makes good sense to assume that the total number of basic words in a given language is always finite. Otherwise, the very idea of a lexicon will stand questioned - an infinite set cannot be enumerated exhaustively within finite space and time. We all have a mental lexicon and we are all familiar with the notion of a dictionary. It is meaningful to list words and their meanings only because the number of words is finite.

All possible phoneme sequences are not meaningful. *The set of words in a given language is a meaningful subset of the set of all possible finite length phoneme sequences*. That is, each word in a language has a definite meaning. Recall the difference between the terms set and sequence. In a set, there are no repetitions and elements are not considered to be in any particular order. A set is simply an unordered collection of elements. A sequence, on the other hand, is an ordered collection. Also, there can be repetitions in a sequence. *Words are sequences of phonemes, not sets of phonemes.*

How many phoneme sequences are possible? Let us say a language has 50 phonemes. Then there are 50 phoneme sequences of length one. There are 50 times 50 or 2500 phoneme sequences of length two. There are 50 times 50 times 50 or 50^3 phoneme sequences of length

3, 50^4 phoneme sequences of length 4, and so on. *Words are always finite in length but this does not mean that there is any arbitrary limit on the length of words in a given language*. One may be curious to know the longest word in a given language or the longest word that has actually occurred in a given corpus. Imposing this length limit, one can even calculate the total number of possible phoneme sequences. You should expect this number to be an extremely big number. This exercise is interesting but it is theoretically not acceptable to impose arbitrary length restrictions because languages often have productive word formation processes that involve endless looping of some kind. The intention of introducing loops is not to go on building longer and longer word forms endlessly but to capture some simple generalization in word structure. Given all this, the set of all possible finite length phoneme sequences should be taken as infinite. Out of this infinite set, we need to extract a finite subset of meaningful phoneme sequences and call them the words of the language in question. How exactly do we do this?

There are two kinds of infinite sets, called countable and uncountable. Countable sets are those in which a one-to-one correspondence can be established between the members of that set and the members of the set of positive integers. This implies that the members of a countably infinite set can be ordered in some particular order and we can talk of the first item, second item and so on. In the case of infinite sets that are not countable, this is impossible. The set of real numbers is not countable. The set of real numbers and the set of whole numbers are both infinite. Yet the set of real numbers is somehow bigger - it includes all the whole numbers but the set of whole numbers does not include all of the real numbers. Understanding these mathematical concepts is very important both for NLP professionals and linguists.

The set of all possible phoneme sequences is a countably infinite set. Countably infinite sets can be generated by a simple mechanical procedure. We can generate all phoneme sequences of length one, then all phoneme sequences of length two, and so on. Phoneme sequences of a given length are simply the combinations and permutations. To illustrate this idea, let us say we have only three symbols called a, b and c. The set of all strings is {a, b, c, aa, ab, ac, ba, bb, bc, ca, cb, cc, aaa, aab, aac, aba,...}

The set of words in a given language is a finite subset of the set of all possible finite length phoneme sequences in that language. We need a precise way of defining this subset. A set can be defined either by listing its elements or by using a rule. For example, { apple, mango, orange, grape, banana } is a set of fruits defined by explicitly enlisting the elements of this set. We can define the set of prime numbers by giving a rule: { *x* || *x is a positive integer, x is not divisible by any integer except 1 and x* }. This is to be read as "the set of all x such that x is positive integer and x is not divisible by any integer other than by 1 and the same number x".

Now, one option we have is to simply list all the valid words in a given language and assert that only what is listed in this list, called the lexicon, is a valid word, nothing else is. This is not a very good option. As language keeps changing, new words come in and some words may go into the oblivion. A fixed list is too rigid. Also, words may have substantial internal structure and many words are best generated by a morphology component. Simply listing all words may be unintelligent and impracticable too. Further, simply making an arbitrary list is theoretically not a satisfactory solution. Who decides what is a word and what is not? On what basis? Language is not designed by any one person or any one central authority. Language is a common wealth of a community. People may not be able to precisely define a word but people do have the general capacity to say what is a valid word and what is not. Therefore, arbitrary listing without any basis is not an acceptable solution. If at all we wish to make a list of all valid words in a given language, the only way we can decide which candidate strings to include and which ones to exclude is to go by meaning. There is no other way.

If we choose the set builder notation to define the set of words, we will have to say { *x* || *x is a finite sequence of phonemes and x has meaning*}. Thus, whether we try to define the set of valid words in a language by enlisting all the words or by giving a rule, the only way we can decide which items to include in the set is to go by meaning. *There is no other way to define words. Words have to be defined based on meaning, not on any other consideration.* This leads to the fundamental question: *what exactly is meaning*?

3.3 Meaning of Meaning

Volumes have been written on this topic but here let us be focused, short and precise. Our immediate concern here is about language, its structure and meaning. We are exploring the mind only to understand the nature and abode of meanings. Interested readers may go through the book entitled "Freedom" by the same author for a more detailed discussion on the nature of the human mind (Murthy, 2012) in relation to language and linguistics.

Whatever we experience in our life, the mind creates some impressions of it and stores them. When I was a child, my father or somebody else showed me a dog and said 'dog'. On seeing the dog, I had created some mental impression about that object. And I learned to associate the uttered phoneme sequence 'dog' to that particular object. Next time I saw a dog in my life, I could recollect the stored impression and match it with the corresponding sound pattern and I could also say 'dog'. *Meanings are impressions stored in the mind*. The impression may include the physical attributes such as shape, size, colour and texture, movements such as wagging of the tail, but it is not restricted to physical attributes. I may find the dog scary or I may find the wagging of the tail amusing and all this is part of the meaning of the word 'dog' for me. The standard technical term for this is *vaasanaa* but for the sake of simplicity, here we have chosen to use the term 'impression'.

It is not necessary that different people get exactly the same impressions under similar experiences. Thus the meaning of the word 'dog' may be different for different people. A cow is a friendly, peace-loving, kind-hearted, intelligent, divine, holy animal for somebody but for another person it may only signify so many litres of milk or so many kilograms of beef. The cow itself is neither holy or unholy, it is all there in our minds. There is no gender bias in any word or linguistic expression, if at all there is any bias, it is there in our minds. *Words do not have meanings, we attribute meanings to words*.

Meanings can also change with time. As we gain more experience in life, our mental impressions can change and so we have a new meaning for the same word. If the same word means different things to different people, how can we communicate at all? That seven and a half Billion people are conducting their daily business using language shows that meanings cannot be totally disjoint across people. Meanings must be largely common, at least at the the gross, physical or material level, otherwise language would simply not work. Nevertheless, individuals can and do have their own subtle impressions that can vary significantly from person to person.

We can directly perceive physical objects, as also certain kinds of actions. These are percepts, as opposed to concepts. It is easy to understand how these perceptual experiences lead to mental impressions. What about abstract concepts? Abstract objects such as beauty and anger, and mental actions such as feeling and knowing can also be conceived by the mind in a similar manner. First comes experience. If we have already experienced anger in our life and the word 'anger' is used to describe this experience in a suitable situation, we can relate our experience to the corresponding linguistic expression. The word 'anger' is thus created in our mind. What if a small child comes across, say, the word 'romance'? The child does not as yet have any experience of romance. How does the child understand the word? Taking clues from the context in which the word is used, the child imagines a meaning and this becomes the meaning of the word for that child, right or wrong as it may be. Later in life, newer experiences can add, refine or even completely change the meaning of the same word. The meanings of all words are not equally sharp and crystal clear in our minds. Some words we understand well and others we only have a vague idea of. The same word can have different meanings to different people at different times. This bridge between experience and expression at the level of words is called the mental lexicon. For the sake of practical convenience, we often say that a lexicon stores words and their meanings, without worrying about the difference between meanings and meaning representations.

Note how we are deliberately avoiding the contextual influences. For example, the word 'door' may mean different things in usage, once it may refer to the door of a house and another time it may refer to the door of a car. Car-door and house-door have significant differences. Even if we restrict to car-doors, there can be significant differences from situation to situation. Here, we are intentionally abstracting ourselves away from contextual and referential influences. A door is a door is a door and the common mental impression we have is the lexical meaning of the word door. When this word is put to use, it gets coloured in various ways but all that is not part of the lexical meaning of the word.

Everybody knows the word 'dog' but we have great difficulty in precisely defining what a dog is. Likewise, people experience great difficulty in defining simple objects in everyday experience such as a chair or a cup. Let us say we define a chair as an object for people to sit on. We can also sit on a stool or a sofa or on the floor or on the branch of a tree if we wish. Therefore, this definition is too vague and too general. If we say a chair is for one person to sit on, we have excluded a bench and a sofa but what about a stool? If we say a chair is a piece of furniture with four legs, a back rest and hand rests, we must realize that not all chairs have four legs or arm rests. There are so many kinds of chairs, new types keep coming in and it is not easy to give a very precise definition. What about the reclining seats we see in a bus? Are they chairs? An ideal definition should set chairs apart from all other things in the Universe. The definition should not apply to anything other than a chair and it must apply to all chairs. It is quite hard to give a precise definition of a chair. If this is the situation with regard to simple physical objects, what about abstract notions such as love, peace or freedom? What about actions? What exactly is walking or bending or eating? The meanings given in dictionaries are only rough indicators, not exact definitions. It is very difficult, well, impossible, to precisely define the meaning of words. The reason is, meanings reside inside our minds, meanings are non-symbolic, they cannot be represented exactly using any kind of symbolism. Words, on the other hand, whether spoken or written, are symbolic expressions. We can only attempt to represent the non-symbolic meanings using symbols but this is only an attempt, we can never succeed fully. Further, it is at best a crude and limited approximation. There are no words in this world to express exactly what we feel inside ourselves. Words are discrete and finite, meanings may not be.

Nevertheless, we need to try hard and give as precise definitions as possible, of all terms used in any serious scientific endeavour, including, of course, linguistics and language technology. For example, it may be sufficient in a given scenario, to define a chair as a piece of furniture with a back rest meant for one person to sit. This is not the only possible definition nor can we prove that this is the best but we may convince ourselves that this definition is good enough.

When we see a picture, our mind perceives the picture as a whole, it creates a mental impression. We can never express the whole of this internal impression precisely and completely in any language. No description of a picture equals the perception of the picture per se. Otherwise, it should have been possible for a painter to create an exact replica of another painting merely based on a verbal description given by a careful observer. This would not be possible even if we have an ideal painter with no limitations. This is impossible because we just cannot express in words what all our mind feels. *We can never say exactly what we feel, we can never say all that we feel. This is the hard limitation of language.*

Yet words are much more precise than pictures. A 'house' is just a house but a picture of a house has so many attributes depicted explicitly. A picture of a house is either a small house or a big one, it has either a flat roof or a sloping roof, it has so and so kind of doors and windows, it is oriented in so and so direction and so on. It is impossible to draw a picture of a house that just indicates a house and not any particular kind of house. *A picture is worth one thousand words. Therefore, one word is much more precise than a picture. Words allow us to say what we want, not more, not less. Pictures and other possible representations are not good enough. That is why we use language as the primary means of communication and thinking.*

A computer has no *mind* at all and so there can never be any meaning inside a computer. However, it is possible to store and recall expressions of meanings. An attempt can be made to express meanings using words. These words themselves have no meaning but they have the capacity to trigger meanings in the minds of the human listeners. In fact dictionaries do exactly this. If we agree to work with verbal expressions of meanings rather than with meanings per se, we can store and recall meanings using a computer too. Then, the computational lexicon would simply be a table mapping phoneme sequences to meaning representations. Note that this table can be used bi-directionally – we can get meanings from given words and we can locate words to convey given meanings. Remember that a computational lexicon is entirely symbolic, it works with meaning representations, not with meanings per se.

3.4 Prosodic Requirements

Meanings are impressions stored in the mind. First comes experience. Experience leads to impressions which are stored in the mind. We can then recall and give expression to these stored impressions. Meaning is all about linking experience and expression through the bridge called language. Let us see how exactly this happens. When somebody says 'dog', we hear three phonemes in a sequence, one after the other. As soon as we hear the phoneme /d/, no particular meaning is triggered in the mind. The listener's mind temporarily stores the phoneme and waits to hear more. Next the vowel sound is heard. Even at this point in time, no meaning is triggered. The mind saves the sequence of the two phonemes heard so far and waits to hear more. Once the last phoneme /g/ is heard, the stored sequence immediately triggers the meaning of a dog and the word is perceived. It is possible that parts of a word are other valid words. Yet the mind may not notice the partial possible words. This happens because there is prosody in speech. Pitch, duration, stress, intonation, and other supra-segmental features indicate word boundaries. All this information is lost in writing. When we hear the word 'understand', the mind is unlikely even to notice the part 'under' as a separate entity. Similarly, when we hear 'had been running', we hear only one word, not three. The three parts need to spoken together, without any gap or discontinuity. You cannot say 'had', give a long pause, and then 'been running'. Nor can we insert any other arbitrary word in between. You can of course say 'had always been running' but you cannot insert words arbitrarily. Nor can you disturb the natural intonation pattern drastically. Parts of a word are spoken and heard with a good degree of continuity. This is an extremely important idea. This can be called *sannidhi,* borrowing from a similar idea applied at sentence level. In 'I had my breakfast', 'had' is a word by itself. In 'I had been to London', 'had been' is a single word. In 'I had been running', 'had been running' is one single word. Think of meanings, think of spoken utterances, do not think in terms of the written form.

Sometimes linguists try to bring in arguments that are logically untenable. For example, they argue that the word 'had' can be 'moved' to form sentences such as 'had he been running?'. Since 'had' has been moved, it must be a word in its own right. This argument is self defeating. Even before you prove that 'had' is a word, you assume that it is word and like other words, this word has moved, and so it is a word. Movement applies to words. Before you establish the word-status of a candidate item, you cannot even think of movement, for, movement applies only to words. The 'had' in the interrogative sentence does not have the same meaning as the 'had' as part of 'had been running' in the assertive sentence. In the interrogative sentence, it is this item which indicates that the sentence is interrogative. In the assertive sentence, the same item has no such tendency. The meaning of a word is fixed, it is based purely on the sequence of phonemes it is made up of, not on anything else. If there is difference in meaning, it cannot be the same word.

4. Universal Word Classes

4.1 Nouns and Verbs:

Nouns are words that denote things. Here the term 'thing' is used in a very broad sense, it includes inanimate objects, living beings such as trees, animals and human beings, place, abstract concepts such as numbers, anger and love, etc. Common nouns and proper nouns are both nouns. Verbs are words that denote action. We can perceive things and actions and create mental impressions of these. Thus, nouns and verbs are words. Nouns and verbs are universal word classes, every language needs to talk of things and what can be done with them. Nouns and verbs have lexical meaning, that is, there is a direct relationship or mapping between the phoneme sequences and stored mental impressions. Nouns and verbs can be understood even when heard in isolation. They have independent meaning of their own. They mean the same thing in all contexts – a dog is a dog is a dog, right?

Nouns are delimited by space. Thus we can talk about spatial attributes such as size, shape, location and orientation. Abstract nouns are delimited in the space of possibilities. For example, anger is delimited by the space of possible emotions. Nouns refer to things and things can be named and described. Verbs, on the other hand, are delimited by time. Verbs indicate actions and actions can have a starting time and an ending time. Actions imply changes taking place over time. That is why we can talk of tense and aspect for verbs. When we say verbs denote actions, it is not necessary that there is any perceivable change taking place. In the sentence 'there is a tree in front of my house', 'is' is a verb, a form of the root 'be'. The existence of the tree can have a beginning and an end. The word 'be' can have tense, so it is a verb. Therefore, stative verbs are also verbs exactly like action verbs and there is no need for a separate definition. Some changes are drastic and patently visible while others may be slow and imperceptible (growing, for example). Yet verbs always indicate action and changes taking place over time. Thus it is easy to recognize and distinguish between verbs and nouns, the two most important and universal word classes.

4.2 Pronouns:

Nouns and verbs have independent, fixed lexical meanings. Pronouns are not words of the same kind. You did not get a specific mental impression when you heard the word 'he' for the first time in your life and you do not simply store and recall the same fixed mental impression every time you hear the word. Pronouns are variables that can stand in place of a noun. The human mind deals with pronouns quite differently from the way it deals with nouns and verbs. Pronouns are not stored in the lexicon, they are stored separately. On hearing a pronoun, we need to dereference it to find out whom or what it refers to. In order to facilitate this reference resolution process, pronouns are usually marked with a number of grammatical features such as gender, number and person. After a pronoun is dereferenced, it stands for some person or object, that is, some noun. The meaning of this noun is of course found in the lexicon. Note that a pronoun can refer to, or stand in place of, not only simple nouns but entire noun phrases. Thus, you can talk of 'the big banyan tree next to the temple' and then refer to this whole thing by saying 'it'. In other words, pronouns can stand for individual nouns as also for higher level grammatical objects constructed from several lexical words and grammatical connectives etc. according to rules of syntax. Hence pronouns should not be simply mixed up with other words and treated on par with them. If pronouns were simply place holders for nouns, they should have had similar grammatical properties, but they do not. You can say 'the boy', 'that boy', 'tall boy', 'third boy' but you cannot say 'the he', 'that he', 'tall he', 'third he' by simply replacing 'boy' with 'he'. Pronouns are not words, just like nouns and verbs. Pronouns are not part of the lexicon, they are part of the grammar. Incidentally, pronouns are also quite irregular in morphology in many languages.

4.3 The Next Tier: Adjectives and Adverbs

Consider the phrase 'the big tree'. Firstly, the meanings of words such as 'big' are relative. A big pencil may be much smaller than a small house. Bigness can be relative to the object being described and it can also be relative to the observer. A cup may be small for a human being but quite big for an ant. When we hear the word 'big' in isolation, our mind does not get any definite, fixed meaning. We understand 'big' only in relation to some object which is said to be big. When we hear the phrase 'a big tree', we hear a sequence of phonemes corresponding to the left to right reading of this phrase. Upon hearing the part 'big', we do recognize it as an adjective, but its meaning becomes fully clear only after hearing the whole phrase. This is true of all adjectives. Secondly, attributes do not have independent existence. The colour of an object depends upon the lighting conditions etc. and hence not an innate and inalienable property of the object. What properties an object can have is defined by what kinds of observations the observer can make. Attributes are imposed by the observer on some object, they have no independent existence of their own. It is possible to have an object without attributes but attributes without an object is impossible. See (Murthy, 2012) for a deeper and more elaborate discussion on this very fundamental idea.

We superimpose attributes on an object and use adjectives to talk of such superimposed attributes. Attributes have no independent existence, they have existence only in relation to some object on which they are superimposed. A car can be red or green but without an object that can have colour, red or green makes no sense. You cannot see or perceive 'red', you can only perceive some object which is red in colour. The mind can conceptualize 'redness' but redness is an abstract noun, not an adjective. There is no meaning for the term 'red' apart from the perception of redness imposed on some real or imaginary object. The term 'big' has no meaning of its own. We do not store specific impressions in the mind on first experience of adjectives and we do not recall the same impressions later. Instead, we perceive objects with certain attributes, we do not perceive the attributes separately. The mental impressions created will be of objects inclusive of the attributes, not of the attributes independently. Adjectives are therefore not lexical words on par with nouns and verbs.

Adverbs of manner are similar - we understand 'slowly' only in relation to some action that is said to go on slowly. Adjectives and adverbs are not pure lexical words like nouns and verbs. However, adjectives and adverbs are semantically motivated classes and so we may accept them as universal word classes, albeit at a dependent level. If we agree to this, then we must consider 'the big tree' as two words, not one, although we are talking of only one thing, one object here. We shall say more about 'the' later.

Words such as 'big' and 'red' may be used to describe the properties or attributes of objects but not all adjectives are used to describe nouns. Adjectives are also used to specify, select, identify, or point to an object. When used predicatively, adjectives do describe objects but when used attributively, they have other functions too. In 'the third bench', 'third' is not a property of 'bench'. It is used only to specify which of the several benches in the universe of discourse we are talking about. Similarly, in 'three books', 'three' is not an attribute of 'books', we are instead creating a new object - a set consisting of three books. In 'the book', the so called determiner 'the' suggests that we should focus our attention on that particular book we have just now been talking about. Thus, 'the' is just a feature marker, it only indicates definiteness, it has no lexical content of its own, it cannot be considered to be a word by itself. Definiteness is a discourse level feature and we must understand this as such. In 'a book', we may be either talking of 'one' book or any indefinite book. In 'that book', the so called demonstrative 'that' is used to deictically select, identify, and point-to the particular book we wish to talk about. In 'my book' and 'my father's book', 'my' and 'my father's' may not be attributes of the book as such. There are several kinds of words, with several different functions but all of them are used to select the objects we wish to talk about. In this broad sense, all of them can be grouped together and called adjectives. Anything which has a syntactic relation of some kind with a noun, except verb-noun relations, is an adjective. Note that only adjectives that describe the properties of objects are allowed to be used predicatively. We can say the 'book is big' but we cannot say 'the book is that', or 'the book is three' or 'the book is third'.

In examples such as 'the poor envy the rich', adjectives are used in place of the nouns they qualify. Predicative adjectives can be interpreted as nouns and in many languages including the Dravidian languages, they take nominal forms too. In Sanskrit, adjectives are always understood as equivalent to the nouns they qualify, and adjectives take gender, number and case exactly like nouns. Grammatically speaking, adjectives are non-different from nouns in Sanskrit. Adjectives do not form a separate class of words in Sanskrit. In other languages, adjectives may be grammatically quite different from nouns and hence a separate class may be a legitimate requirement but it must be clear that adjectives have only a secondary or dependent meaning.

Manner adverbs can be treated as a universal word class on par with adjectives but in practice all kinds of items are grouped under the head adverb. Some modify verbs, some modify adjectives, some modify the whole predication. Adverbs call for a much more closer and deeper study. As a general rule, all words which act as modifiers in a syntactic relationship with other words, excluding adjectives, and excluding the noun-verb relations, can be called adverbs.

4.4 Non-Words

In modern linguistic tradition, a distinction is usually made between content words and function words. This terminology is problematic. If we talk of content words and function words, we must first define a word and only then partition the set of words into two mutually disjoint subsets termed content words and function words respectively. Since words can only defined as meaning bearing phoneme sequences, functions words are not words at all and so this terminology is logically flawed. Nevertheless, bowing down to tradition and conventions, we may continue to use the terms 'content words' and 'function words'. Otherwise, we will have to invent some new terminology to deal with these non-words and new terminology can also be confusing until firmly established.

Content words have lexical content or meaning whereas function words have more of a grammatical function. Function words do not have clear lexical meaning of their own but they are not completely nonsensical sequences of phonemes either. Lexical or content words are part of the lexicon and function words are really not part of the mental lexicon, they are in fact part of the grammar.

Distinguishing between lexical and functional words is important. Function words are not part of the lexicon. They are not translated directly, they may even be completely lost in translation. Usually function words have little or no morphology. Function words have a greater role in syntax. Function words vary widely across languages or language families. There are no articles in Kannada, nor are there any prepositions. Function words typically act as connectives of various kinds. Some tokens are merely bundles of features – these tokens should be treated as parts of words they are associated with, not as independent entities. Free morphemes are those morphemes that have independent lexical meaning and can so occur independently. Here occurring independently does not mean written as a separate unit surrounded by spaces, it only means independence in terms of meaning. Bound morphemes have little or no lexical content and they only add grammatical features to other words. Bound morphemes should be considered as part of the words they relate to, not as words in their own right. For example, prepositions and post-positions are bound morphemes, not words. We will have more to say on these later.

After taking out bound morphemes, some more entities may be left over, which may have lexical content in some small degree. Should we consider them as lexical words or as function words? If the main role played by such entities is relating two words in a sentence, they should be treated as function words and included in the grammar, not in the lexicon. One reason we have accepted adjectives and manner adverbs to be lexical words is that they do not serve to grammatically relate two other words, nor can we treat them as grammatical features of other words, so they cannot be treated as function words.

Interjections can be pregnant with meaning. However, the meanings are often too abstract and loaded, and heavily dependent on the situational context. Also, they are rarely connected with other words in a sentence through syntactic relations. Interjections are more like sentences in their own right. For these reasons, interjections are usually left out of the rest of analysis.

Articles such as 'a', 'an' and 'the' are not lexical words – no specific meaning is triggered in the mind when these words are heard in isolation. The articles 'a' and 'an' indicate indefiniteness – a grammatical feature of the noun they qualify. The articles 'a' and 'an' can also mean 'one', and that is completely different. 'The' signifies definiteness. Definiteness is a discourse level feature, it indicates that the particular thing we are talking about is already clear in the mind of the listener. We say 'the sun' because there is only one sun and so there is no confusion. We say 'the boy' to talk about a particular boy, which particular boy we are talking about being clear from the discourse context.

The best way to deal with terms that indicate morphological, syntactic or discourse level features of other words is to attach them to the appropriate words. Thus 'an apple' is one word, not two. 'from the school' is one single word. If take this approach, we will be left with only lexical words and grammatical connectives like 'and', 'but', 'if' and 'unless'. The structures of sentences in all human languages will come much closer to one another and to sentential meaning. A whole lot of ambiguity and confusion that exists in NLP and linguistics will melt away.

4.5 Prepositions

Perhaps prepositions are the most confused of all items in modern linguistics. The term itself simply signifies that these items are placed before nouns, nothing more. Modern generative linguistics has come to assign a major role to prepositions in syntax. In reality, prepositions are not words at all.

Prepositions are actually degenerate forms. It is often said that prepositions relate two nouns. 'The book on the table' is an example of this. The preposition 'on' is said to relate the nouns 'book' and 'table'. In reality, this is a degenerate form of 'the book which is/was on the table'. A verb is essential. In fact in many languages including Dravidian, we cannot express this without explicitly including a suitable verb. There is a verb, the subject of this verb is the book, table is the place. The preposition 'on' merely indicates the location role the word 'table' takes with respect to the verb. Prepositions are very much like case markers, they are merely grammatical features, not words in their own right. Confusions arise because prepositions are degenerate forms. Consider the following sentences:

- He saw the accident. He fainted.
- He saw the accident and he fainted.
- After he saw the accident, he fainted.
- After seeing the accident, he fainted.
- On seeing the accident, he fainted.
- At the sight of the accident, he fainted.

In the first case, we have two separate sentences. In the second, we have a single coordinate sentence. In the third, we have a main clause and a subordinate clause. Here 'after' is acting as a subordinating conjunction. In the next example, the gerundial form 'seeing' is used, the verb 'see' is becoming a bit nouny, and 'after' is more like a preposition. This is substantiated by trying a different preposition, 'on', which is never a subordinating conjunction. In the last example, the verb 'see' has fully become the noun 'sight' and 'at' is clearly a preposition. These examples human languages do show mixtures and gradations. Prepositions are actually degenerate forms of case markers and subordinating conjunctions.

Certain prepositions such as 'in' and 'on' may have a bit of meaning but when heard in isolation, prepositions do not trigger clear and fixed impressions in the mind and as such, we cannot accept prepositions to be lexical words. Prepositions are grammatical units, they relate a noun and a verb, they are grammatical features, not words. It is best to understand prepositions as features attached to nouns, very much like the case markers in Dravidian. The so called post-positions need to be treated exactly the same way.

'The book' is one word, not two. 'The big book' is not one word, although it signifies a single object or thing because we have accepted adjectives as words in their own right. As we have already seen, 'the' is not a word, it only signifies the definiteness feature of the noun. Thus there are only two words here, not three. What about 'the book on the table'? 'on the table' is the location where 'the book' 'is /was'. 'The boy from Delhi' means 'the boy who has come from Delhi' or something like that, and in that case 'from Delhi' is the source location from where the 'the boy' has 'come'. Similarly 'from Delhi' is one word, not two. 'The boy from Delhi' has two words with an implied verb.

We are interested here in unearthing the underlying universal properties of all human languages. Prepositions are certainly not universal, many languages do not have prepositions at all. The term Pre-position only signifies placement before some other entity and there is no direct reference to sound or meaning. Grammatical features, whether at morphological or syntactic or discourse level, should not be confused with words.

4.6 Grammatical Properties

Generally speaking, nouns and pronouns take number and case, verbs take tense and aspect. Adjectives modify nouns. All other modifiers are called adverbs. Fine, but there can be word forms that show mixed behaviour. Gerunds are verbs that have become nouns. Like nouns they take case but like verbs they may also take objects, be modified by adverbs etc. Similarly, infinitives are verbs with some nouny properties. Verbs can also become adjectival. Such derived word forms have been a major source of confusion. A verbal noun is neither a pure verb nor a pure noun and it would be wrong to label it either as verb or as noun. We believe that the best way to treat such derived words is to record the complete process of derivation at all levels including tagging, chunking and syntactic parsing.

Morphological and syntactic properties are both important. Syntactically, nouns and pronouns can act as subjects and objects. Nouns and pronouns can be heads of noun-phrases. However, nouns and pronouns differ in significant ways. Pronouns cannot be modified by adjectives, nouns can be. Nouns can also modify other nouns, pronouns do not. Pronouns are not modified by demonstratives or determiners, nouns can be. Thus, although we generally say pronouns stand in place of a noun, you cannot simply replace a noun with a suitable pronoun in a given sentence.

Semantics, morphology and syntax do not always go together. In terms of semantics, the so called spatio-temporal nouns are adverbs of place or time. Morphologically, they can take some case markers although not all. Syntactically, some of them can act as subjects, although they rarely do. What should we do when we face such confusing indicators from semantics, morphology and syntax?

From the point of view of universality, it is always better to give maximum importance to semantics and consider syntax and morphological considerations as secondary level. However, computational systems in NLP often start from the bottom, do morphological analysis of the given surface text first and work upwards towards syntax and hence semantics. Computers do not directly understand meanings and we do not know how to start from semantics. In such situations, the general tendency is to start with morphological considerations, make adjustments if required to prepare for the syntactic level and hence to semantic level. For example, adverbs of place and time may be considered nouns as they take cases.

Compounds and phrasal constructions also pose confusions. Orthographic conventions vary a lot. Is the Kannada 'snaana maaDu' (lit. bath do) one word or two? How about 'uuTa maaDu' (lit. meals do) and 'kelasa maaDu' (lit. work do)? Here 'maaDu' (do) is a verbalizer for constructing noun-verb compound verbs. We need to apply linguistic criteria to decide. Can we insert other words in between? Can we move one component away? Can we modify one part? We cannot say 'maaDida snaana' (the bath which was done/taken), 'maaDida uuTa' (the meals which was done/taken) but 'maaDida kelasa' (the work which was done) is fine. We cannot describe 'snaana' (bath) using adjectives (including demonstratives, quantifiers, ordinals, possessives, etc.) but we can describe 'kelasa' (work). Although 'maaDu' is a transitive verb, we cannot say 'avaniMda maaDalpaTTa snaana' (the bath which was done/taken by him). Thus, 'kelasa maaDu' is two words while 'snaana maaDu' and 'uuTa maaDu' are single words - as they indicate a single atomic kriyaa.

Compounds are made up of two or more words but they should be treated as single words at the level of the lexicon, morphology and syntax because they carry single lexical meaning. The meaning of compounds is non-compositional, we need a whole sentence involving other words to describe their meaning fully. Thus a 'water meter' is a meter used for measuring the flow of water, a 'water pump' is a pump used for pumping water and a 'cast iron pump' is a pump made up of cast iron. Irrespective of whether compounds are written together, hyphenated or orthographically separated by white spaces, we must consider them as single words. On the other hand, sandhi (phonetic conflation) between two or more words involves no changes in meaning or grammatical properties and so the component words should be treated as separate words. Thus, 'avaniiga' in Kannada needs to be broken into 'avanu' (he) and 'iiga' (now).

Here we have only given some general directions for asking linguistic questions and thereby deciding the candidature of a given item for word-hood. For a more detailed discussion on how to define grammatical categories and sub-categories, readers may go through our papers on the topic (Murthy and Srinivasu, 2012, 2013)

5. Word Defined

In order to bring about a complete closure to our discussion on words, word classes, subcategories and their grammatical properties in full detail, we also need to take a fresh and detailed look at syntax. We need to start by asking what exactly is a sentence. We define a sentence as a set of words, inter-related to one another through syntactic relations, forming a single connected graph. (We have used the term 'set' here, in order to make it clear that strict ordering of words is not an essential aspect of sentence, not to imply that there cannot be repeated occurrence of a given word in a sentence. A sentence becomes a 'set' after words are indexed, distinguishing between dog_1 and dog_2, for example.) Syntax is basically relating words in a sentence to one another through syntactic relations. Which relations are syntactic and which are discourse relations is a critical question here. Syntactic relations connect two words while discourse relations connect sentences. Thus, we can precisely define a sentence only after we have defined the complete set of syntactic relations. Identifying a universal, semantically motivated set of syntactic relations is the big challenge. As a general observation, we can remark here that there are mainly two kinds of syntactic relations: a) noun verb relations specifying the role played by the noun in the action indicated by the verb b) modifier-modified relations such as between an adjective and a noun, between two verbs, etc. Then we need to precisely define each of these syntactic relations and thematic roles. Then for each language, we need to work out the mappings from surface indicators to the thematic roles. All this is beyond the scope of the present paper. However, keeping these ideas in mind, *we can now define a word as a sequence of phonemes bearing a* *clear lexical meaning (even when used in isolation) and bearing certain kinds of syntactic relations with other words when used in a sentence*.

Lexical words form part of the mental lexicon. In contrast, the so called function words are phoneme sequences with a clearly identifiable grammatical function such as connecting two parts of a sentence. These latter units are not part of the lexicon, they are part of the syntax. Only lexical words must enter the lexicon, nothing else. Lexical words are defined starting from meanings and so there should not be multiple entries for a given meaning. Barring synonyms, various spellings found in usage cannot be considered as separate entries in a lexicon. Morphology should deal with words, one word at a time. Neither part of a word nor more than one word can be entertained. Thus 'had been dancing' and 'jaa rahaa thaa' are single, atomic words, they need to be sent to morph and analyzed accordingly. Tags should be assigned to words and only to words. As we have already seen, there are no such things as auxiliary verbs, we should not accept such arbitrary tags. Many a time, what people call a phrase is actually a single word. Shallow parsing, chunking etc. need to be seen in this light. Similarly, terms such as 'local word grouping' and 'multi-word expression' need to be looked at afresh. We believe that all these are not at all necessary. Syntax should deal only with grammatical relations between words in a sentence, nothing else. 'The' is not a word, and so syntax should not worry about relating 'the' to 'book' in the construction 'the book'. Similarly, prepositions should be used to connect nouns to verbs, not for anything else. Thus, our view of word has far reaching implications for all levels of linguistic analysis and language processing by man or machine. Note again that these ideas have nothing to do with the written form. When viewed from this perspective, languages of the world will lose many of their apparent differences and we will start seeing the underlying universal properties of human languages. There will be much closer correspondence between the words in equivalent sentences in different languages. Words will not be as ambiguous as we may be thinking now. Translation will become so much simpler. Language independent technologies will begin to rise.

Nouns, verbs, adjectives and adverbs are found in most languages of the world and these four can be taken as universal word classes. Only these four kinds of words enter the lexicon. Since these word classes have been defined by taking resort to meanings, it is expected that as a general rule, there will be one-to-one correspondence between these words across human languages, subject to influences such as culture, history and geography. It is not true that English sentences are much longer than Telugu sentences, in terms of number of words. This cannot be the case. Also, since words are defined starting from meanings, it is not right to say one word has many meanings. If there are many meanings, there are as many words. Spellings may or may not be different. In some cases, pronunciation also may or may not be different and in this case ambiguity is real. It is unlikely that most words have many meanings in any language, if that is true, it will become very difficult to work with such a language. As a rule, one word has one meaning. Sense ambiguities are an exception rather than a rule. 'Word Sense Disambiguation' is not as hard as you think it is. Identifying words can be a challenge, though.

A good test for word-hood is whether the item can stand on its own as a reply to a suitable question in a suitable discourse situation. Nouns and verbs can stand on their own. For example, 'Q: what is that?', 'A: tree', 'Q: what are you doing?', 'A: eating', etc. are acceptable. Adjectives and adverbs have no independent existence and that is why they generally require the help of the noun or verb being modified. For example, 'Q: which book?', 'A: big book' is OK but we cannot simply answer the question by saying 'big'. This query-response test establishes the fact that prepositions, post-positions, particles, conjunctions, etc. are not words. Needless to add that meta level questions such as which is the two letter English word that begins with 't' and ends with 'o' are not allowed here.

Further, since we have defined words starting from meanings, a word must have the same meaning in all usages. If it becomes difficult to establish a clear and fixed meaning to a token, perhaps it is not a word at all. One word, one meaning is the rule, ambiguity is an exception. Therefore, we must expect more or less word to word mappings in translation, alignment etc.

Note that polysemy is the very nature of words. Meanings are in the mind and words are discrete symbols we use as labels for a whole range of meanings. There are many different varieties of trees, small or big, tall or bushy, and we call all of them by the term 'tree'. Cooking may involve using a gas stove or an electric cooker or stones and fire wood and the verb 'cook' includes all these various possibilities. There is no need to distinguish between all these various possibilities at the level of linguistic analysis and NLP.

Note that roots and bases are words as also are the inflected and derived word forms. A word form may have a root/base which gives the lexical meaning and other affixes which give the grammatical part meaning.

5.1 Identifying Words

So far we have tried to answer the question 'what is a word'. The question that still needs to be answered is, how do we identify proper words in a given text? Note that in speech there are anyway no characters or spaces and we are all used to recognizing words based on meanings. All the confusions arise the moment we start working with written texts. As human readers, we can go by meanings and recognize and understand words and sentences in even written texts. When we wish to automatically recognize words using a computing machine, then this question is legitimate, because computers do not understand meanings. Whatever be the situation, let us first understand and accept that the written tokens do not always correspond to words and the disparity can be glaring in many language-script situations. Let us first accept that we cannot take the incorrect path, however hard the correct path may appear to be. We should identify words before we proceed. If required, we can do this with human intervention.

In some languages, notably Sanskrit and the Dravidian languages, the difference between the written units and proper words is not so much. A little bit of careful pre-processing using Regular Expressions or Finite State Automata will do in many cases. When it is really hard to divide sentences into proper words, as in the case of compound versus sandhi (phonetic conflation between two or more words), we can either take care of these within morphology or introduce a bit of post-processing (a kind of morpho-syntactic bridge) after which, we will definitely have the sentence divided properly into words and analyzed accordingly. This is doable. We have actually implemented comprehensive systems including lexicon, morphology, tagging etc. for Telugu and Kannada, thereby supporting our ideas and substantiating our claims. The theoretical foundations we are trying to lay, the architectures and software systems we are developing, including the linguistic data, are known by the label 'the saara system'. Versions of these software systems are available for free download from our website (202.41.85.68) and readers may wish to take a look.

In other languages, including English and Hindi, the units obtained by tokenizing written sentences based on white spaces differ quite a bit from proper words as we have defined in this paper. Nevertheless, simple techniques such as Regular Expressions and Finite State Automata can be used to re-group tokens into proper words. Once this is done, English and Hindi will not be so very different from Telugu or Kannada. Words will become longer and more complex, morphology will become richer, sentences will become smaller (in terms of number of words) and syntax will become so much more similar to what it is in

Kannada and Telugu. In fact all languages in the world will start looking so much more similar to one another at all levels. This is the way to go.

Dravidian languages have a great advantage here and we must try and put our languages in the fore front of linguistic studies and language technologies at the global level.

6. Conclusions

In language technology and NLP, as also to a great extent in modern linguistics, words are being defined as sequences of characters separated by spaces. In this paper, we have argued against this idea and tried to define words in a more universal and language independent way. We have tried to show the merits of taking this view. The goal of this paper is not to give a final answer to the question of what exactly is a word but only to rekindle interest in this most important topic and to assert that we can actually do much better than what we seem to be doing with languages in the world today. Our own work on Telugu and Kannada definitely stand as strong evidence in support of our ideas.

Bibliography

1. Allwood J, Hendrikse A P, & Ahlsén E, *Words and Alternative Basic Units for Linguistic Analysis*. In Henrichsen, P.J. (Ed.) *Linguistic Theory and Raw Sound*. Copenhagen Studies in Language, 40:9-26. Copenhagen: Samfundslitteratur. 2010

2. Baker M.C, *Lexical Categories: Verbs, Nouns and Adjectives*, Cambridge University Press, 2003

3. Booij G.E, *The Grammar of Words: An Introduction to Linguistic Morphology*, Oxford University Press, 2005

4. Fromkin V, Robert Rodman, Nina M. Hyams, *An Introduction to Language*, Thomson/Heinle, 2003

5. Graddol D, Cheshire J, and Swann J, *Describing Language*, Open University Press, 1994

6. Kavi Narayana Murthy, *Natural Language Processing – an Information Access Perspective*, Published by Ess Ess Publications, New Delhi, for Sarada Ranganathan Endowment in Library Science, Bangalore 2006

7. Kavi Narayana Murthy, *Freedom – The Art and Science of Life*, Literary Circle, Jaipur, 2012

8. Kavi Narayana Murthy and Srinivasu Badugu, *Roman Transliteration of Indic Scripts*, Proceedings of the 10th International Conference on Computer Applications, 28-29, February, 2012, University of Computer Studies, Yangon, Myanmar

9. Kavi Narayana Murthy and Srinivasu Badugu, *A New Approach to Tagging in Indian Languages*, Research in Computing Science, Issue 70, 2013, pp 43-54.

10. Larson R.K. And Ryokai K, *Grammar as Science*, MIT Press, 2010

11. Larry Trask, What is a Word? - A Working Paper, Department of Linguistics and English Language, University of Sussex, 2001.

12. Kavi Narayana Murthy, Srinivasu Badugu, *On the Design of a Tag Set for Dravidian Languages*, 40th All India Conference of Dravidian Linguists, 18-20, June, 2012, University of Hyderabad, Hyderabad, India

13. Stekauer P. And Lieber R, *Handbook of Word-Formation*, Springer, 2006