Words and alternative basic units for linguistic analysis

Jens Allwood
SCCIIL Interdisciplinary Center, University of Gothenburg
A. P. Hendrikse,
Department of Linguistics, University of South Africa, Pretoria
Elisabeth Ahlsén
SCCIIL Interdisciplinary Center, University of Gothenburg

Abstract

The paper deals with words and possible alternative to words as basic units in linguistic theory, especially in interlinguistic comparison and corpus linguistics. A number of ways of defining the word are discussed and related to the analysis of linguistic corpora and to interlinguistic comparisons between corpora of spoken interaction. Problems associated with words as the basic units and alternatives to the traditional notion of word as a basis for corpus analysis and linguistic comparisons are presented and discussed.

1. What is a word?

To some extent, there is an unclear view of what counts as a linguistic word, generally, and in different language types. This paper is an attempt to examine various construals of the concept "word", in order to see how "words" might best be made use of as units of linguistic comparison. Using intuition, we might say that a word is a basic linguistic unit that is constituted by a combination of content (meaning) and expression, where the expression can be phonetic, orthographic or gestural (deaf sign language). On closer examination, however, it turns out that the notion "word" can be analyzed and specified in several different ways. Below we will consider the following three main ways of trying to analyze and define what a word is:

(i) Analysis and definitions building on observation and supposed easy discovery
(ii) Analysis and definitions building on manipulability
(iii) Analysis and definitions building on abstraction

2. Analysis and definitions building on observation and supposed easy discovery

We will start by considering analyses and definitions intended to build on observation of linguistic communication. Here the idea is that words are the basic building blocks of linguistic communication, providing combinable units of meaning and external expression that should as such be fairly directly observable and discoverable when inspecting linguistic communication, whether in written, spoken or gestural form. This can especially be seen in the definition of orthographic words given below.

(i) **Orthographic words**

According to Trask (2004) “[a]n orthographic word is a written sequence which has a white space at each end but no white space in the middle”.
This definition of ”orthographic word” is both too wide and too narrow, in relation to other notions of word that intuitively have precedence. For example, the expression ”rail road” has two orthographic words but is intuitively one word. This means that the notion of orthographic word as defined captures too much, i.e. it is too wide => too many words.
But making ”rail road” into two orthographic words is also too narrow => not capturing the word (semantic unit and phonological stress unit, lexeme) that is actually there.

(ii) **Phonological words**

Following Trask again, Trask (2004) defines a phonological word as “a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language”

Unfortunately, there are units other than words that perhaps meet such phonological requirements, for example phonemes, syllables or breath groups. The definition does not tell us how to differentiate these units from each other. The mention of language-specific features does not help, since, these might be different for different languages, for the different units. In addition, when transcribing words, i.e. making them into orthographic words, typical phonetic information that may be used in the identification of phonological words such as stress, tone patterns, pauses (length) are either typically not represented in transcriptions. One reason for this is that such information is not traditionally part of written language, another is that it may reflect that this information is not so easily consciously recognized/observed by phonetically untrained transcribers.

If we consider the relation between orthographic words and phonological words, we first may note that given these two definitions of a word, a consequence is that there is no 1- 1 correspondence between orthographic words, phonological and semantic words. Consider the following examples: *rail road* (2 orthograpic words - 1 phonological word*) or I’m, you’re, won’t* and *ain’t* (1 orthograpic word - 1 phonological word but two semantically motivated words). *New Yor*k (2 orthographic words) vs. *Newfoundland* (1 orthographic word). *New York and Newfoundland*, thus, fairly arbitrarily, have different orthographic status while both probably are single phonological words etc.

(iii) **Gestural words**

Using Trask’s definition of phonological words as a model, we can now define gestural words analogously as as “a piece of gestural communication which behaves as a unit of gesturing according to criteria which vary from language to language”

The relation between orthographic, phonological and semantically motivated and gestural words is more complex, so that 1 – 1 correspondences between the three word forms are not always possible to establish here either. Concerning gestural languages (sign languages), one reason for this is that while written and spoken words can be seen as variants of the same unit in two different expressive modes, gestural words in sign language are units in a new language and not gestural variants of the same word, in the sense that the written and spoken variants of a word are variants.

We can also note that only the definition of orthographic word is operational, i.e. lives up to the desiderata of being both directly observable and discoverable and thus, directly usable as an element in automated information retrieval. Since, as we have seen above, the criteria given for what is a unit of pronunciation or gesturing are not sufficient, these concepts thus remain in need of further specification and clarification.

3. Analysis and definitions building on manipulability

Many linguists have thought that word criteria, based on inherent word features that are supposed to be directly observable are unreliable and need to be supplemented by other criteria. Some widely used such criteria are criteria that in a syntactic mode focus on the unit status of words. Two criteria are often suggested:

(i) Moveability
(ii) Resistance to intrusion and interruption

Both of these criteria have often been used to define the notion ”word”. We will now consider them one by one.

3.1 Moveability

According to this criterion, a word is the smallest element of a sentence that can be moved around without destroying the grammaticality of the sentence.

Thus, the fact that the word *often* in the expression *often he went to the house* can be moved from first to last position as in *he went to the house often*, shows that *often* is a word. A problem with this criterion is that several kinds of units that have not traditionally been considered words can be moved around in a similar fashion. Consider the following examples:

(i) Movement of morpheme

*The **un**faithful wife was masked -> The faithful wife was **un**masked*

Even *if -un* is not usually regarded as a word but as a morpheme, it can be moved around without destroying the grammaticality of the embedding sentence. The meaning is changed, but since the criterion wisely does not demand preservation of meaning, *-un* passes as a moveable unit. If preservation of meaning had been required, even the example given above, using the word *often*, might not qualify (the information structural aspect of meaning is changed). In fact, very few changes of word order do not have an effect on meaning and it would not be a trivial task to say which aspects of meaning do not change when word order is changed. If it be objected that there is no movement here, *-un* is just deleted and affixed to a new unit, a reply to this is that movement can always be analyzed as a combination of deletion and addition and since there is no requirement of preservation of meaning there is no way to rule this type of example out.

(ii) Movement of phrase

By and large you are right -> You are right by and large

The phrase *by and large* is not usually regarded as a word but clearly behaves in a word like fashion, using this criterion. Thus, if we use the criterion of moveability, it seems that morphemes and fixed phrases are somewhat arbitrarily excluded from word status.

Since especially what one might call "lexicalized phrases" are important for our argument, we will give some more examples of expressions of this type that arguably have lexicalized status. The classification and examples are taken from Moon (1998) (cf. also Wray 2002).

1. Different types of "anomalous" collocations

At all, by and large, of course, stay put, thank you, in retrospect, kith and kin, on behalf of someone/something, short shrift, to and fro

at least, a foregone conclusion, in effect, beg the question, in time, curry favour, foot the bill, toe the line

in action, into action, out of action, on show, on display, to a …degree, to a …extent

2. Fomulae

Simple formulae
alive and well, I'm sorry to say, not exactly, pick and choose, you know

Sayings
an eye for an eye, curiouser and curiouser, don't let the bastards grind you down, that's the way the cookie crumbles, home, James, and don't spare the horses

Proverbs
you can't have your cake and eat it, enough is enough, first come first served

Similes
good as gold, as old as the hills, like lambs to the slaughter, live like a king

3. Metaphors

Transparent metaphors
alarm bells ring, behind someone's back, breathe life into something, on (some) one's doorstep, pack one's back

Semi-transparent metaphors
grasp the nettle, on an even keel, the pecking order, throw the towel in, under one's belt

Opaque metaphors
bite the bullet, kick the bucket, over the moon, red herring, shoot the breeze

3.2 Resistance to 'intrusion and interruption'

The second common criterion for word-hood is that words are the largest units which resist 'intrusion and interruption' by the insertion of new material between their constituent parts.

Following this criterion, we can insert the word *not* in the expression *it is uninteresting* and obtain a grammatical expression *it is not uninteresting.* However, we cannot easily insert *not* in the expression *uninteresting* and obtain *un-not-interesting* as a grammatical expression in English. Words but not sentences are resistant to intrusion and interruption.

But again, as with the criterion of moveability, this criterion does not seem to always hold. In some circumstances some words seem to admit some interruptions. Consider the following examples:

Absolutely -> abso-bloody-lutely, underdeveloped -> under-bloody-developed, unmanageable -> un-fucking-manageable etc.

In both cases, i.e. "moveability" and "resistance to intrusion and interruption", the criteria are not sufficient, rather they serve as guidelines that are often but not always useful.

4. Analysis and definitions building on abstraction

The same word can be used several times, e.g. should we count the sequence *horse, horse, horse* as 3 words or 1 word? Since both options seem to be valid, it is useful to distinguish between:

(1) Word tokens and word types, where tokens are concrete instances (in time and space) of a conceptual type, e.g. the sequence *car, car, car* presents 3 tokens of the same word type *car,* where the type is based on abstraction over the tokens.

Both word types and word tokens can be either orthographic, phonological or gestural but, as we have seen above, there is no 1-1 correspondence between orthographic, phonological and gestural word tokens and word types, e.g. *rail road, rail road, rail road* constitutes a sequence of three two word orthographic tokens (corresponding to one two word orthographic type) but at the same time, it also corresponds to three phonological word tokens and one phonological word type.

Two other word related notions based on abstraction appear in the distinction between:

(2) Lexemes and word forms

The expressions *speak, spoke, spoken* are three word forms of one lexeme that can be represented by one of the word forms, for example, by the infinitive "speak". This word form also has other functions like present tense or imperative. However, the form is chosen since it is common to all of the mentioned forms and infinitives are traditionally often chosen to represent root forms for verbs in English and other languages

Taking the type - token distinction as our point of departure, we can say that the lexeme is a sort of super word type, while the other word forms are word sub types. All of the types can

then in turn be related to tokens and any type can be said to represent the set of tokens related to it.

Lexemes and their corresponding word forms can also be related to phonological, orthographic and gestural words but it is not obvious that there can be phonological, orthographic and gestural lexemes and word forms, since lexemes and word forms seem to be more abstract and very much semantically motivated. Should we say that the orthographic word *yes*, the spoken word *yes* and a head nod represent the same lexeme? Or should we say that they are three different lexemes representing the same "sememe" (a term sometimes used for a basic semantic unit)? Should we say for spoken and written variants of the same word, that they represent the same lexeme or that they are different lexemes representing the same sememe? These sememes (or lexemes, depending on what we decide) might then possibly be said to have spoken and written variants that in turn have either phonological or orthographic tokens. As we have seen it is a little harder to claim that gestural words share sememes or lexemes with spoken and written language (at least if they come from sign language), since sign languages are mostly not just signed variants of a given spoken or written language but separate languages, cf. above section 2 (iii).

It is worth noting that the word forms belonging to a lexeme add semantic information to the so called root or stem of the lexeme. In the example above "past time" is added in *spoke*, while something like "past completion" is added in *spoken*. According to the traditional view, word forms can add semantic information as long as the added information does not make the lexeme change its part of speech category. Thus, *speakable* is not normally regarded as word form of *speak* but as a "derived form" making up a new lexeme. If one wanted to have a unit that included derivations and possibly also compound forms, this could be done using labels like "inflectional lexeme", "derivational lexeme" and "compound lexeme" (or possibly "inflected", "derived", "compounded lexeme", depending on whether one wanted an external holistic view or an internal process oriented in-use view).

It is also worth noting that lexemes are biased toward categorematic parts of speech (nouns, verbs and adjectives) and are less useful in relation to syncatagorematic parts of speech (adverbs, prepositions, conjunctions, articles, numerals, interjections) in the sense that the latter categories in most languages don't have inflected forms. They are also biased against moveable inflectional units like a genitive *-s* added to the end of a noun phrase, e.g. the genitive *-s* in *the man in the green car's money*. In general, the concept of lexeme has not been used to identify word forms from the point of view of their inflectional or derivational morphemes, e.g. giving a list of the forms belonging to the lexeme for the derivational suffix *–able* would be to list all the words that have been derived with *–able*, like *readable, edible, constructable* etc..

The results of our brief survey show that the notion of a word can be specified in the following ways:

(i) graphic material between spaces,
(ii) piece of produced speech,
(iii) piece of produced gestured message,
(iv) unit that can be moved without changing grammaticality,
(v) unit that resists intrusion/interruption,

(vi) unit that allows for meaning related abstraction on several levels (tokens, types, word forms, lexemes capturing inflected forms, derived forms or compound forms or even sememes, capturing common semantic content).

We therefore see that the notion of word can be specified in several different ways even in relation to one language like English. What are the consequences of this for linguistic theory? What are the consequences when we want to compare of different languages on measures involving words?

5. Words, corpus analysis and typological comparison

The orthographic word is usually taken as a point of departure for automatic analysis of corpora and has been the basis for several attempts to capture for example the following linguistic features and types of analysis: Size of corpora, word frequencies (Parts of speech frequencies), MLU (Mean length of utterance), Concordances (KeyWordInContext (KWIC)), Grammatical patterns (colligations), Lexical patterns (collocations), Vocabulary richness, Lexical density, Translation alignment

All of these attempts have been made in relation to written language corpora, where they are based on the notion of orthographic word we have discussed above. Since the measures are increasingly used in many ways, where the most important use is probably comparisons of different types (e.g. corpus size, frequence), it means that the problems we have seen concerning orthographic words can skew the results. This problem becomes very serious when we want to compare (or translate between) typologically different languages with different conventions of word formation. To illustrate this point, let us consider some examples of translation and thus of comparison between English and agglutinative languages:

As we have seen above, in analytic languages, the word is a unit that is supposed to be smaller than a phrase or a sentence that both typically consist of a sequence of words. But this account does not hold straightforwardly for all languages. In polysynthetic languages, it can be difficult to draw a distinction between sentences and words. Here is a typical sentence from Yup'ik, an Eskimo language of Alaska:

Kaipiallrulliniuk. 'The two of them were apparently really hungry.'

The sentence consists of a verb stem *kaip-* 'be hungry' followed by a string of suffixes. In effect, the whole sentence is merely a grammatical form of this verb and if we compare the number of words in the two languages, we find that eight English words correspond to one Yup'ik word.

We encounter similar problems if we compare English and Xhosa. Compare the English sentence and its Xhosa equivalent below:

The dog of the neighbour bit me
I-nja yo-mmelwane i-ndi-lum-ile
a/the-dog of-neighbour it-me-bit

In this case, 7 English “words” correspond to 3 Xhosa “words”. Commenting on this lack of correspondence between analytic and agglutinative languages, Hendrikse and Poulos (2006: 260) note,

“There are no free morphemes in the agglutinating African languages in such word categories as noun, verb and adjective. With the exception of a few words, mostly of an adverbial, conjunctive, interjective and feedback nature, all word-forms in Xhosa are morphologically complex constructions.”

The comparative examples clearly show that using quantitative automatic word based measures in these two language types will yield significantly different and even compromised results because they measure different and incomparable entities. This, in turn, fairly clearly shows that comparisons between analytic and agglutinative languages (e.g. English and Xhosa) using quantitative automatic word based measures are highly problematic. In fact, on reflection it should be clear that none of the commonly used word-based features/measures can be used and be expected to deliver a non-biased result, in comparisons between languages, since they are usually based on the orthographic word.

It is therefore very natural to ask the following question: Should the word as a basis for measures of features of language and for linguistic comparisons be replaced by another unit? Alternatively, one might also consider whether any or all of the various specifications of the notion of a word, suggested above, instead could be used in making such comparisons.

6. Are there alternatives to the traditional notion of word as a basis for corpus analysis and linguistic comparison?

At present it is not at all clear that there is an alternative to the traditional word notion as a basis for corpus analysis. Maybe instead we should use different units for different purposes or become very much more specific about what properties of words we are relying on in our analysis.

However, some general observations can be made. In spoken/multimodal corpora, the contributions (utterances and/or gestures (gestures are here defined as all communicative body movements, e.g. including facial gestures)) communicators make to the process of communication, often consisting of single words, gesture or phrases rather than sentences (compare example below), can often be seen as the basic units of communication.

A: One return ticket to Montagu
B: <nods> ten Rand
A: here
B: thanks

Sentences and words are very much connected with written language. Words occur in both written and spoken language but sentences (in the sense of subject + predicate verb constructions) are more rare. Multimodal corpora therefore open the challenge of basing automatic analysis on features not present in written non-interactive communication. Examples of such features are utterances, gestures and prosody and the relationship these features have with linguistic features also shared with written language, cf. Allwood (2008).

In order to continue our discussion of alternatives, we should first consider whether any or all of the specifications of the notion of a word we have noted above can be used in linguistically related analysis, i.e. (i) graphic material between spaces, (ii) piece of produced speech, (iii) piece of produced gestured message, (iv) unit that can be moved without changing grammaticality, (v) unit that resists intrusion/interruption, (vi) unit that allows for meaning related abstraction on several levels (tokens, types, word forms, lexemes (capturing inflected forms, derived forms or compound forms) or sememe).

Probably all of them can be used for different specific purposes. However, it would be important to actually relate them to such purposes. Some of them in addition need considerable further specification before they can be used. This probably holds for all units except "graphic material between spaces". Some examples might be the following: (i) graphic material between spaces – this is a quick way to find words in most Indo-European languages, (ii) piece of produced speech, (iii) piece of produced gestured message – both (ii) and (iii) would be interesting if we wanted to do automatic analysis of psychologically real units, but require more empirical and conceptual specification before they can be used in this way, (iv) unit that can be moved without changing grammaticality, (v) unit that resists intrusion/interruption – (iv) and (v) could be useful in word processing programs, (vi) unit that allows for meaning related abstraction on several levels (tokens, types, word forms, lexemes (capturing inflected forms, derived forms or compound forms) or sememes) – this type of unit could be useful in linguistic comparison. One suggestion here is to take root and stem elements/morphemes of words rather than words or word forms as the point of contact between languages. To this can be added the assumption that semantically, root and stem elements are primary activators of "meaning potentials" which when used are "contextually determined" through inflection, derivation, compounding, syntax or other mostly extra-linguistic contextual information (cf. Allwood (1999 and 2003). This would then be one of the semantics based options for comparison of languages to be mentioned below (and also one of the ways to interpret the notion of a sememe mentioned above).

Let us now take a look at some of the alternatives to an analysis based on words that could be considered:

(i) Syllables, phonological words; recognition of these units is so far non-automatic and the units are perhaps not very often the most natural substitutes for words, given that they have no semantic motivation. They might have a use as a measure of phonological complexity or a use in Speech synthesis if we would like to create a program that generates speech with metric structures. They also have a role in speech recognition with an output in syllabic script like Japanese hiragana or katakana. Automatic recognition of syllables or phonological words would thus have a greater role to play as multimodal corpora become more common.

(ii) Lexemes; lexemes are problematic in agglutinative languages, since they have no clear independent root forms and they have very many derived and inflected forms. Lexemes in agglutinative languages, in this way, come to resemble what we have called inflectional + derivational + compound lexemes above.

(iii) Morphemes; this is probably the second best alternative but discontinuous morphemes and vowel alternation morphemes are a problem in many languages (e.g. semitic languages), resembling the problems we have already noted for words. In such cases and even in others,

morphological analysis is not so easy to do automatically. This means that making use of morphemes would require a fairly large work effort in order to annotate the corpus manually.

(iv) Semantic coding; this would be optimal but is very laborious and so far non-automatic. In addition, the exact nature of the semantic units that should serve as a basis for linguistic comparison is in no way obvious, e.g should it be semantic primitives like in classical generative semantics [male, female young etc], cf. McCawley 1968 or English basic words, supposed to be universal, e.g. Wierzbicka (1996) or more vague semantic field like structures, e.g. the field of epistemic verbs, weather verbs or evaluative adjectives etc allowing for slightly different semantic structuring in different languages? If we base our analysis, on primitives, we need a set of semantic/pragmatic features that would be as universal as possible or at least cross linguistically applicable. A set of narrowly defined semantic primitives will not suffice, since language use is always thoroughly pragmatic. If we base our analysis on some type of field structure, the field probably has to become an activity field in the sense of a script for an activity, e.g. the restaurant script or the bus-driver passenger script, etc., cf. Schank and Abelson 1977, to make room for pragmatic meaning.

Let us now consider whether there are other alternatives not based on words that could help us solve the problems noted above. Below are some suggestions for how we could use other alternatives than word based specifications to obtain information that has often been obtained through the use of the orthographic word:

Size of corpora: words => characters. In the example discussed above (English *The dog of the neighbour bit me* corresponding to Xhosa *I-nja yo-mmelwane i-ndi-lum-ile*), this would lead to 25 characters for English and 24 characters for Xhosa. This would be better than the 7 words vs. 3 words obtained using the orthographic word. However characters are also problematic since many orthographies represent single sounds by bi-graphs or even tri-graphs and sound combinations by single graphs. Corpus size is often interesting in getting a picture of what lies behind reported relative measures or frequencies. It is also a convenient way of getting an idea of the size of the linguistic resources of a given language.

Word frequencies (POS frequencies) => morpheme frequencies; this suggestion would give better results than words. In the example above, 8 morphemes in English vs. 9 morphemes in Xhosa, but is problematic since it requires morphological analysis, which is not directly available in the same convenient way as words are, identified through word space.

MLU => character/morpheme per utterance; Rather than using words, mean length of utterance could just as well or better be measured through characters or morphemes. However the problems noted above in relation to characters and morphemes would still be valid.

Concordances (KWIC) => word/morpheme/utterance in context; Useful concordances can be produced using morphemes or utterances. In this case, perhaps more as complementary types of analysis than as replacements of words, since they would all give slightly different information (e.g. a concordance of utterances can give you pragmatic, functional information, a concordance of words can give you information about polysemy and a concordance based on inflectional morphemes could be the basis for a more thorough semantic/pragmatic analysis of the meaning of the morpheme.

Grammatical patterns (colligations): such patterns can be found in terms of utterances, morphemes or words; this is clearly an interesting alternative not yet significantly explored. However, it is also a difficult task and usually requires manual coding

Lexical patterns (collocations): utterance, morpheme, word; As with concordances, this should perhaps be seen more as complementary types of analysis than as replacements of words, since the patterns in this case also would all give slightly different information. Attempts have been made to find such units automatically, e.g. by trying to find all units that have a higher frequency than is to be expected from the frequencies of its constituent words. However, again work remains to be done and it is likely that manual coding would still be necessary.

Vocabulary richness => utterance/lexeme/morpheme/word richness; this is a type of analysis that tells you how many different word types there are in a given corpus and since all the units mentioned, i.e. (utterance, lexeme, morpheme, word) can be used for type – token abstraction, they can all be used to give different type of richness measures (or if so desired, the opposite stereotypicality measures where those texts are more stereotypical that show a greater repeated reuse of the same words and constructions, cf. Allwood and Sjöström (2001), where stereotypicality measures were given for texts coming from the Swedish ministry of education)

Lexical density => morphemic density; this is a type of analysis which relates categorematic words (nouns, verbs and adjectives) to syncategorematic (function) words (adverbs, prepositions, conjunctions, articles, numerals, interjections) as a ratio, so it would be hard to replace the word. Possibly root- and stem-morphemes that in context have been classified according to part of speech could be used.

Translation alignment: multiple alignment; Translation alignment is used as a basis for automatic translation and involves building up a store of already made translations (a so called translation memory) based on an alignment between the words of two different languages. This clearly becomes very problematic if 3 words in one of the languages correspond to 7 in the other, like in the examples given above of Xhosa and English. Simple word based alignment could be replaced by multiple alignments based not only on words but also on e. g. morphemes, phrases, words, punctuation marks (for written language), utterances and utterance position (for spoken language). Such multiple alignments would be likely to increase the level of accuracy of the translation memory significantly.

Finally, we would like to point out that neither the types of analysis proposed on the basis of a refined analysis of the notion of a word, nor the units and types of analysis proposed above as alternatives to using words do sufficient justice to all of the information that might be present in a corpus. This is especially relevant in relation to multimodal/spoken language corpora consisting not only of transcriptions but also of digital audio and video recorded material, cf Allwood (2008). In such corpora, features like utterances, gestures and prosody and the relationship these features have with other linguistic features must be taken into account. One very simple initial way of doing this is to align the transcriptions and other annotations made with the audio and video recordings. Another perhaps simpler suggestion would be to be to base more analysis on utterances (in the sense of a complete contribution from a speaker until the next speaker takes over (types and tokens)), and position of a construction in an utterance. Since, however, utterance length varies considerably with activity, comparisons based on number of utterances would then also have to take this feature into account. This would be

useful, especially in interactive spoken language and result in a linguistic structural analysis closer to the actual structure of interaction.

7. Concluding remarks

In this paper, we have argued that it is problematic to assume the ”word” as a basic linguistic unit for the retrieval of various standardized types of information from corpora. We have also argued that the notion of ”word” is polysemous and has been associated with at least 7 different meanings.

The notion of an ”orthographic word” is especially troublesome in being both too narrow and too wide even intra-linguistically. To use it as a basis for inter-linguistic automatic comparisons gives very skewed results. Finally, we have noted that, at present, it is unclear what notions and constructs should replace the ”word”. We have made some suggestions and noted that probably several notions will be needed for different purposes.

References

Allwood, J., 1999. ‘Semantics as Meaning Determination with Semantic Epistemic Operations’ in J. Allwood and P. Gärdenfors (eds.) *Cognitive Semantics*, pp. 1-18. Amsterdam: Benjamins.

Allwood, J. and S. Sjöström. 2001. ‘Ordförrådet i läroplanerna’, *Gothenburg Papers in Theoretical Linguistics* S25, 2001, Department of Linguistics, University of Gothenburg.

Allwood, J., 2003. ‘Meaning Potential and Context. Some Consequences for the Analysis of Variation in Meaning’ in H. Cuyckens, R. Dirven, and J. R. Taylor (eds.) *Cognitive Approaches to Lexical Semantics*, pp.29-65. Berlin: Mouton de Gruyter.

Allwood, J.,2008.’Multimodal Corpora’ in A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. pp. 207-225. Berlin: Mouton de Gruyter.

Hendrikse, A.P. and G. Poulos. 2006. ‘Tagging an agglutinating language: A new look at word categories in the Southern African indigenous languages’, *Language Matters* 37 (2), pp. 246-266.

McCawley, J. D., 1968. ‘The role of semantics in a grammar’ in E. Bach, and R. Harms (eds.) *Universals in Linguistic Theory*, pp. 124-169. New York: Holt, Rinehart and Winston

Moon, R., 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Alderley: Clarendon Press.

Schank, R.C. and R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale: Hillsdale Publishers.

Trask, R.L., 2004. ’What is a Word?’ *Working Papers* 11, Department of Linguistics and English Language, University of Sussex.

Wierzbicka, A., 1996. *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Wray, A., 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.