INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1222)
Available online at: https://www.ijariit.com

Measures of lexical diversity

Vijaya
vijuciefl@gmail.com
The English and Foreign Languages University, Lucknow, Uttar Pradesh

ABSTRACT

The "natural" and the "formal" acquisition of a language by a child are best seen as two ends of a continuum. The acquisition of languages involves a combination of formal instruction and natural acquisition. This paper captures an overview of select prominent tools used for measuring vocabulary diversity across first and second language contexts. The tools include type-token ratios (TTR) and the more recent automated measure of VOCD.

Keywords: *Second Language, Vocabulary Measures, Void, Type-Token Ratio, Lexical Diversity*

The "natural" and the "formal" acquisition of a language by a child are best seen as two ends of a continuum. In many cases the acquisition of so-called "foreign" languages involves "a combination of formal instruction and natural acquisition" (Schreuder & Weltens, 1993, p. 3). The outcome of second or foreign language instruction is bilingualism, and research on second language vocabulary imperceptibly shades off into research on the nature of the bilingual lexicon.

...[T]he L2 learner may also, like the first language learner, have more or less organized knowledge of paradigmatic relations between words that share features of meaning and/or form, and of syntagmatic relationships between words that co-occur in language use. (Harley, 1995: 3)

1. RESEARCH ON SECOND LANGUAGE VOCABULARY

Lexically oriented second language research has witnessed a steady growth since the 1980s (Meara, 1996; Harley, 1995a, 1995b). The prominent themes of research have been the *processes in second language vocabulary comprehension, learning and use* and *the nature of growth of the second language lexicon.* Meara (1996:36), describing the limitations of classical research on vocabulary acquisition, notes that "a lot of current research is making the same mistakes." Many studies treat vocabulary acquisition as memorizing lists of paired words as arbitrary signs. We may see the tradition of comparing the learning of cognate and non-cognate words as originating in similar presuppositions regarding words as arbitrary signs: with the arbitrariness apparently reduced in the case of cognates. Such studies also look only at the learning of a tightly-controlled subset of words in experimental settings. As Meara (op. cit.) points out, "learning a set of 20-40 words may pose some difficulties for short-term memory, but seen from a long-term perspective, and in comparison with the number of words a fluent speaker needs to know …, such numbers are basically trivial."

Crucially, the larger vocabularies that underlie creative language use in comprehension or production in real time are not lists of words with their sounds and shapes and meanings, but are words organized in ways being investigated by studies of the mental lexicon. Finally, many such studies employ only the weakest criterion of word or meaning recognition as evidence of acquisition, rather than the ability to put a word to active use.

Knowing a word is not an all-or-nothing affair. Even in the first language, the "complete learning of a word's meaning is a gradual process probably extending over years of time in which the word is encountered repeatedly in contexts" (Nagy, Herman and Anderson, 1985: 236). This position on vocabulary acquisition, referred to as "incidental vocabulary acquisition", the "uninstructed learning position", or the "default hypothesis position", opposes the view that vocabulary growth is the result of explicit vocabulary instruction (Beck, Perfetti, and McKeown,1982). The incidental view of vocabulary acquisition, which maintains that exposure to the target language through reading plays a key role in improving vocabulary, has been extended to the second language context by Elley and Mangubhai (1983), and Krashen (1989).

1.1 Vocabulary Measures

Measures of vocabulary whether in first language or second language usually fall into two categories: "breadth", or "depth" measures. Breadth measures, the more frequently employed of the two, provide a rough estimate of the overall size of the vocabulary. Depth measures are concerned with certain aspects of meaning, word associations or use (Verhallen and Schoonen, 1993; and Wesche and Paribakht, 1996). Depth refers to how well a word is known.

Vocabulary breadth measures rely on frequency lists. Lack of knowledge of frequent words reflect a smaller vocabulary. Some commonly used frequency lists in English (cited in Wesche and Parbakht, 1996) are: The General Service List (West,1953), Thorndike and Lorge's Teacher Word Book (1944), University Word List (Xue and Nation, 1984), and Kucera and Francis' Word List (1967). Dictionaries have also been used for selection of target vocabulary (e.g., every *n*th word in the dictionary). Vocabulary size is taken as:

$$\frac{No.of\ coorect\ answer\ \times Total\ words\ in\ dictionary}{No.of\ items\ in\ test}$$

Nation (1993) proposes procedures for sampling words from dictionaries for such tasks. The commonly employed test methods include multiple-choice (Nation, 1990: 75), checklist (Read, 1988; and Nation, 1990), C-test (Klein-Braley, 1985; Singleton & Little, 1991), dictation (Nation, 1990), judgement of "real" words (Harlech-Jones, 1983), and matching words with translations or definitions (Nation, 1990).

Vocabulary depth measures do not seem to be as frequently employed as breadth measures. Various aspects of word knowledge have been described: "different meanings, appropriate uses, syntactic properties, underlying forms and derivations, association networks with other words and experience, and connotations" (Wesche & Paribakht, 1996:26). There also seem to be different levels of word knowledge and use capabilities. They vary from initial recognition of the graphic or oral form of a word, to understanding of its common meanings in context, to the ability to produce a word rapidly and accurately in a range of contexts.
Nation (1990) proposed four dimensions of word knowledge: form of a word (oral, written), position (grammatical, collocations), function (frequency, appropriateness), and meaning (concepts and associations). Some of the depth measures employed are: the 'word associates' test developed by Read (1994), quoted in Wesche and Paribakht (1996), and Wesche & Paribakht's (1996) VKS-Vocabulary Knowledge Scale.

Other vocabulary measures include the lexical decision task and indices of lexical diversity or richness. In a lexical decision task, the time taken to recognize a string of letters as a word or a non-word is treated as a measure of lexical access, or "look up." As we are interested in capturing the appearance of nouns and verbs and other open and closed class categories in our study, we shall now focus on the measures used for studying lexical diversity or lexical richness in the spontaneous speech data.

1.2 Lexical Diversity

Lexical diversity is a term used to refer to vocabulary breadth in production, as against vocabulary depth. For example, a speech sample containing 20 "tokens" of two "types" is considered more diverse than a speech sample with 40 tokens of one type. In order to measure lexical diversity or richness, a number of indices have been used with written and oral texts. The most popular of them all seems to be the type/token ratio (TTR)[i] (Arnaud, 1984, cited in Wesche & Paribakht, 1996:24). Vermeer (2000) provides an integrated summary of measures used over the years in measuring lexical diversity in spontaneous speech data.[ii]

1.2.1 Lexical Diversity and Noun Advantage: Among the measures described by Vermeer (2000), the lexical diversity measures used in studying early vocabularies have been (i) percentages, (ii) TTR, and (iii) distribution-type of "flat"/"steep" distribution. The earliest of the studies reporting a noun advantage started with analyzing the data in terms of percentages (Stern, 1924, Nelson, 1973, and Gentner, 1982). The measure appearing next chronologically in the language acquisition literature seems to be the TTR (Fletcher, 1985 and Lieven, 1978).

1.2.1.1 Type/Token Ratios: TTR seems to be the most frequently applied of the methods described by Vermeer (2000). TTR is calculated as follows:

$$\text{Type/token ratio} = \frac{\text{types: number of different words (of a given kind)}}{tokens: total\ number\ of\ words (of\ that\ kind)}$$

The measure TTR is still in use. More recently, it is being used as a predictor of foreign language learning aptitude and attainment (Skehan, 1986 cited in Richards, 1987).

According to Richards (1987), TTRs have frequently failed as a measure of lexical diversity. "…TTRs calculated from a large number of tokens will generally be lower than those calculated from a smaller number" (Richards, 1987: 203). For instance, a child with a type frequency of 5 and a token frequency of 50 would yield the same ratio of 1/10 as another child with a type frequency of 50 and a token frequency of 500. (See Chapter Four, section 4.2.1 for an example showing this effect in our second language data.)

1.2.1.2 "Flat" and "Steep" Distributions: The TTR measure has been improved upon in two ways. One suggestion is from Sandhofer et al. (2000) made in the context of a controversy regarding the role of input in the first language acquisition. Sandhofer et al. (2000) added a new dimension to the controversy: how should input be measured? They suggest that if nouns and verbs present different kinds of learning problems to begin with, as assumed by the natural partitions hypothesis, then the kinds of input for nouns

and verbs cannot be the same and cannot be measured by the same metric. They make their argument by adding another dimension of description to the input, in addition to "types" and "tokens." This is the dimension of "semantic category." Their argument is that if most of the nouns in the input and output belong to the same semantic category, while the verbs belong to different semantic categories, then the kinds of input for the learning of nouns and verbs will also show differences in optimality.

The common nouns that children encounter early may, as names for concrete things, present a common semantic, conceptual, and syntactic structure. In particular, most early nouns may be count nouns that name concrete things in shape-based categories….In contrast, the verbs young children encounter early do not present equivalent semantic and syntactic structures. For example, the two common verbs *put* and *look* describe highly dissimilar actions and relations between arguments. Understanding the meaning of *put* would seem to provide little insight into the meaning of *look.*
(op. cit. p.p. 564)

Sandhofer et al. (2000) propose to supplement the metric of frequency with that of distribution, i.e. the type-to-token ratio of nouns and verbs and its relationship to the semantic categorization. They motivate the measures of "steep distribution" and "flat distribution" in a first language input study. A "steep distribution" of words refers to tokens of a few particular words (i.e. types), while a "flat distribution" refers to many tokens distributed across many different words (i.e. types).

However, most measures including the TTR present researchers with the problem of a variation in the lexical diversity along with a change in the sample size. Richards (1987) raises a theoretical problem with the measure. The type/token ratio of a 500-word text tends to be lower than that of a 200-word text. This reflects the fact that as we produce more language we tend to use fewer and fewer of those words which have not been already used earlier. TTRs fail to capture the variety of usage in more advanced vocabularies. This shows that TTR may not be a very reliable measure of lexical diversity.

1.3 Lexical Diversity and the Measure 'D'

In view of the problems involved in using TTR, Richards (1998, cited in McKee, Malvern & Richards, 2000) has suggested a new measure of lexical diversity, the measure, 'D'. D is calculated using the program *vocd.* The program automates measurement from transcripts prepared on the basis of the CHAT format. D is expressed in terms of a ratio very much like the TTR, but unlike TTR it is not effected by the variations in sample size.

The new measure is calculated by the programme's first randomly sampling words from the transcripts. A curve for type/token ratio against tokens is made. "Then the software finds the best fit between this empirical curve and theoretical curves calculated from the model by adjusting the value of a parameter" (McKee, Malvern & Richards, 2000:323). The parameter, D, is argued to be a reliable measure of vocabulary diversity. The advantages, as pointed out by McKee et al., of using D are: 1) it is not a function of the number of words in the sample; 2) it uses all the data available; and 3) it is more informative because instead of a single value of type/token ratio, it represents how the type/token ratio varies over a range of token size for each speaker.

A higher D shows a better lexical diversity. "...[V]alues have been found to range from D=5 for a five-year old language impaired child to D=120 for a sample of academic writing" (Malvern and Richards, 2002: 90). The program yields types, tokens as well as TTRs along with two D values: D optimum value and D optimum average. The program provides different ways of calculating lexical diversity. D can be calculated 1) excluding the inflected forms, that is, each inflected form constitutes a single type; 2) considering regular inflections as one type; and 3) reducing irregular forms (Richards (2004) refers to this as "D (root forms)").

2. CONCLUSION

In this paper, we capture the measures of types, tokens and type-to-token ratios along with the measure 'D' for analyzing lexical diversity in the spontaneous speech data in first and second language settings. As the language develops, the counting of mere tokens or number of words gives way to the counting of 'types', or the number of different words used by the child, as a measure of the diversity of its vocabulary. The 'type-token ratio' is a classical 'breadth' measure of vocabulary diversity. This has given way to more recent sophisticated vocabulary 'breadth' measure being used in recent studies on vocabulary, 'D' (VOCD: McKee, Malvern and Richards, 2000).

3. REFERENCES

[1] Arnaud, P. J. L. (1984). *The lexical richness of written productions and the validity of vocabulary tests. Practice and problems in language testing* (pp 14-28). Papers from the 7th international symposium on language testing, Colchester, UK.
[2] Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74,* 506-521.
[3] Elley, W. B., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly, 19,* 53-67.
[4] Fletcher, P. (1985). *A child's learning of English.* Oxford: Blackwell.
[5] Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development*: *Vol. 2. Language thought and culture* (pp. 301-334). Hillsdale, NJ: Lawrence Erlbaum.
[6] Harlech-Jones, B. (1983). ESL proficiency and a word frequency count. *ELT Journal*
[7] Harley, B. (1995b) Introduction: The lexicon in language research. In B. Harley, *Lexical issues in language learning* (pp.1-27). Philadelphia: John Benjamins.
[8] Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal, 73,* 440-464.

[9] Kucera, H., & Francis, W. N. (1967). *A computational analysis of present day American English.* Providence: Brown University.
[10] Lieven, E. V. M. (1978). Conversations between mothers and young children: Individual differences and their possible implication for the study of child language learning. In N. Waterson & C. Snow (Eds.), *The development of communication*. Chichester: Wiley.
[11] Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19(1),* 85-104.
[12] McKee, G., Malvern, D., & Richards, B. J. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15(3),* 323-337.
[13] McWhinney, B. (1995). *The CHILDES Project: Tools for analyzing talk*, 2nd edition. Hillsdale,N.J.: Erlbaum.
[14] Meara, P. (1996). The classical research in L2 vocabulary acquisition. In G. Anderman and M. Rogers (Eds.), *Words, words, words: Translator and the language learner* (pp. 27-39). Clevedon: Multilingual Matters.
[15] Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20,* 233-253.
[16] Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.
[17] Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential but rarely followed procedures. *Language Testing, 10,* 27-40.
[18] Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs for the Society for Research in Child Development, 38.*
[19] Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19(2),* 12-25.
[20] Richards, B. J. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*, 201-209.
[21] Sandhofer, C.M., Smith, L.B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning. *Journal of Child Language, 27,* 561-585.
[22] Skehan, P. (1986). *Early lexical development and the prediction of foreign language learning success.* Paper presented at the CILT/ESRC conference on second language acquisition, Windsor, Berkshire.
[23] Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In G. Hall & S. Waxman (Eds.), *Weaving a Lexicon* (pp. 257-293). Cambridge, MA: MIT Press.
[24] Stern, W. (1924). *Psychology of early childhood upto the sixth year of age*. New York: Holt.
[25] Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30, 000 words*. New York: Columbia University Teacher's College.
[26] Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics, 14(4),* 344-365.
[27] Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1 / L2 acqisition and frequency of input. *Applied Psycholinguistics, 22(2),* 217-234.
[28] Vijaya (2001). *Input and acquisition of nouns and verbs in the first and second languages*. Unpublished M.Phil.dissertation, Central Institute of English and Foreign Languages, Hyderabad, India.
[29] Vijaya (2001). *Input and acquisition of nouns and verbs in the first and second languages*. Unpublished M.Phil.dissertation, Central Institute of English and Foreign Languages, Hyderabad, India.
[30] Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review, 53(1),* 13-40.
[31] Xue, G. Y., & Nation, P. (1984). A university word list. *Language Learning and Communication, 3*, 215-219.

Notes to Chapter Two

[i] "In theory, Type/Token Ratio (TTR) weights range of vocabulary for size of speech sample" (Richards, 1986, p. 201).

[ii] **Table 2: Vocabulary Measures from Vermeer (2000)**

Measure	Label	Formula
Total number of words	Tokens	N
Number of different words	Types	V
Number of dictionary entries	Lemmas	
Number of types occurring only once	Hapaxes	
Type/token ratio	TTR	V/N
Corrected TTR	TTR ©	$V / 2\sqrt{N}$
Indice de Richesse	gUIRAUD	$V / \sqrt{N}$
Index of Herdan	Log TTR	log V / log N
Uber Index	Uber	$(\log N)^2 / (\log N - \log V)$
Theoretical vocabulary	Menard	