A Survey on Lexical Richness Measures

Petkar H.J.

Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalata,Wardha, Maharashtra.

***Abstract*- Lexical richness measures are used widely across the different disciplines and is part of applied linguistics research. This paper reviews the various lexical measures based on the ratio of type-tokens. Paper also discusses the variants of lexical richness measures based on the square root variants and other indices. Four sections in this paper review the approaches, techniques and work related to lexical richness measures.**

***Keywords*- lexical richness, type token ratio, applied linguistics**

I. Introduction

Lexical richness is one of the dimension of quantitative analysis of the lexical structure. It is also known as vocabulary richness or lexical diversity, lexical sophistication, linguistic variability. There are many measures that have been used for calculating lexical richness. One of the popular measure used for calculating lexical richness is based on the ratio of different words (Types) to the total number of words (Tokens) known as the Type-Token Ratio (TTR).Sometimes lexical richness is measured by the square root variants of TTR and other indices. Quantitative Analysis of the text with the help of TTR and its variant indicate the organization of texts and behavior of the lexical frequency, quantitative text analysis, authorship attribution and quantitative stylistics.

This paper is organized in four sections, The first section introduces the approaches of lexical richness along with the significance. Second section discusses the various existing lexical richness approaches and techniques which are TTR-based measures. In third section detailed review of literature is conducted to explore the work related to lexical richness and it's applied aspects in various discipline. Fourth section presents the concluding remark.

II. Approaches for Lexical Richness

Lexical words are words such as nouns, verbs, adjectives, and adverbs that convey meaning in a text. There are variety of texts in which reoccurrences of the lexical words are observed repeatedly , on the other hand some text avoid the repetitions and use different vocabulary to convey the same meaning, e.g. word important can be replaced with 'principal', 'key', 'supreme' etc. according to context. Greater lexical richness enhances the ability to comprehend the information neatly and shape the way we express our thoughts. It shows the progression enhancement of lexical richness in young people.

Lexical Richness is multi-dimensional feature of written or spoken Language. Lexically rich written script uses the technical terms and jargons along with uncommon words that allow to express their meaning in precise and sophisticated manner.

First approach to measure the lexical richness is capturing the Lexical richness by means of measure of an indices, second one is capturing the unfolding of the vocabulary by the curve whose representation are

Herdan, Tuldava, Kohler, Galle and several Russian Scholars.(Emmirich Kelih 2020). Third one is by using the empirical distribution of words (types) occurring x-times (tokens) and deriving the theoretical distribution based upon combinatorial considerations. Fourth one is based upon stochastic processes resulting in distributions which are pure mathematical approach by Brainerd, Gani, Haight, McNeil, Simon and others (Wimmer and Altmann, 1999).

III. Review of Previous Work

Lexical richness is a matter of concern in variety of fields. Lexical diversity, lexical richness and lexical variability / variety are the alternative terms (Jarvis 2013).Over the last 70 years various lexical richness measures have been developed which is part of applied linguistics research. Most frequently used text-length dependent measure is TTR. (Baayen 2001, Malvern et al 2004).

TTR measure of lexical richness and the several of transformations have been widely criticised and proved to be more or less depend on text length. (Baayen Tweedie 1998, Malvern et al 2004, Kettunen, 2014) .

So it is necessary to review and develop the text length independent measure (kubat- milicka 2013), Covington M.A. &Mc fall J.D.(2010) .

To overcome the length of the text issue in TTR, several suggestions are posed which include the standardizing the length of the text samples. They are mean segmented TTR ,moving average TTR (Covington and Mcfall 2010), Measure of Lexical Diversity(Mc Carthy and Jarvis 2010) . HD -D (Mc Carthy and Jarvis 2007). TTR (Kapantzoglou *et. al.* 2019)), Mean Word Frequency (MWF) (Tweedie & Baayen,1998).

The simple TTR along with its variant such as moving average TTR(MATTR) & other models are used for approximation of morphological complexity of language (kettunen 2014).

There are several categories in which measures of lexical richness can be divided(Torruella et. al 2013) First category is based on the relationship of number of terms and running word in the text (n) known as TTR(Type-token Ratio) which constitute TTR, RTTR(Root type-token Ratio), CTTR(corrected type-token ratio)

Second class of TTR is conceptualized on the basis of logarithmic function. This logarithmic function grows in such a way so as to adopt better to the behavior of the relation that exist between the terms and total number of words such as Herdan, Summer, Mass, Dugest

Third Category of TTR constitutes more complex calculations. MSTTR (mean segmental type token Ratio) in which text is divided into equal size segments of number of words (eg. 100 words per segment). For each segment, the TTR is calculated and using an arithmetic mean of the TTR for each segment MSTTR is obtained.

MTLD (Measure of textual Lexical diversity) (Mccarthy 2005)- The first phase of this measure is similar to that of MSTTR. This phase is segment size dependent. Here Mc Carthy (2005) proposed that instead of depending on the segment size, it is controlled by TTR value ie when it reaches a value of 0.72 upon the operation of segment extension,it will end its operation .At the end of the text the MTLD = , calculation is applied, where L is text length in number of words and n is the number of segment.

In HD-D measure unlike MSTTR and MTLD, it does not uses sequential segments but sample made up

of words selected at random. The length is set as 42 words which can be picked from anywhere from the text. " Given the huge number of possible samples, the average is not calculated directly but via the calculation of probabilities using the hyper geometric probability distribution.

The recent work by (Malvern et al 2004) is developed for child language acquisition is the D- measure which model the rate at which the new words are introduced in increasingly longer text samples, by way of curve fitting procedures which uses on parameter, parameter D.

As describe above in the review more complex transformations are MTLD (measure of Textual Lexical diversity), MTLD-MA (moving average measure of textual lexical diversity), HDD (Hypergeometric Distribution D) or vocd-D argued to be independent of text size (McCarthy & Jarvis 2010)

Determining linguistic complexity of text in terms of lexical richness is one of first basic steps in natural language processing. Linguists use the concept of vocabulary richness mostly in authorship and genre analysis (Kubát, M., & Milicka, J. 2013).

Most of the studies conducted to find the lexical richness of academic writings(Djiwandono 2016). Djiwandono(2016) compared the academic essay of students and teacher and investigate the lexical richness in academic writings. .

Laufer and Nations (1995) derived the lexical frequency profile, which is based on the assumption that frequent words are easier than infrequent Words . This view is supported by Meara(2005).

In Corpus linguistics, quantitative methodology is used to demonstrate the sociovariational distribution of Lexical richness. Various studies demonstrate the lexical richness(Torruella et. al 2013), Malven 2004, 2004 Baayen 2001 Mc Carthy, 2005). All the work calculate the lexical richness by means of statistical formulae.

The longer text implies lower TTR. ie. there is less chance to use the variety of types in whole text. To overcome this limitations the whole text is divided into equal size chunks. To illustrate the influence of the topic on lexical richness measure, an analysis per part of speech is performed, (Baayen 2001) .This experiement tested the difference between TTR's for noun.

The lexical diversity measures are used in the different field of linguists viz. authorship detection (Layton 2012) forensics linguistic (De vel et al 2001), stylistics (Toolan 2009) and increansingly in foreign Language teaching and Learning. Lexical richness measures were used to calculate the proficiency level of child or student and illustrated by comparing their lexical richness with an external reference point. Lexical richness is calculated by the sociolinguistic point of view in (Brindle 2016). In the field of mathematical linguistics and quantitative stylistics, Baayen (2001) introduced word frequency distributions which is characterized by very large number of rare words and results in statistical phenomenon which is mean frequency.

Conclusion

Lexical richness is relevant is the field of Natural Language Processing, Computational Linguistics, Psycholinguistics, Lexicology, Stylometrics, Quantitative Linguistics. Along with these fields measures of lexical richness is used in a wide range of linguistics, applied linguistics and educational research including the development of literacy, forensic linguistics, authorship studies, stylistics, language impairment, child language development, , studies of schizophrenia and many other areas in particular.

Lexical richness is also dominant measure of how language learner acquire and deploy their vocabulary. Paper reviewed various TTR and TTR-based measures which are proposed and demonstrated for the purpose of lexical richness.. It is also found that the calculation for lexical richness, lexical diversity, is widely studied for the languages majorly English, Duch, French(Paula Dissan 2018) on various corpus.

References

Tweedie, , & RH Baayen. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*.

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Netherlands. https://doi.org/10.1007/978-94-010-0844-0

Brindle, A. (2016). *The language of hate: A corpus linguistic analysis of white supremacist language*. Routledge, Taylor & Francis Group.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type– token ratio(Mattr). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, *30*(4), 55–64. https://doi.org/10.1145/604264.604272

Djiwandono, P. I. (2016). Lexical richness in academic papers: A comparison between students' and lecturers' essays. *Indonesian Journal of Applied Linguistics*, *5*(2), 209. https://doi.org/10.17509/ijal.v5i2.1345

Jarvis, S. (2013). Capturing the diversity in lexical diversity: Lexical diversity. *Language Learning*, *63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Kapantzoglou, M., Fergadiotis, G., & Auza Buenavides, A. (2019). Psychometric evaluation of lexical diversity indices in spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *62*(1), 70–83. https://doi.org/10.1044/2018_JSLHR-L-18-0110

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, *21*(3), 223–245. https://doi.org/10.1080/09296174.2014.911506

Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339–349. https://doi.org/10.1080/09296174.2013.830552 Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307 Layton, R., Watters, P., & Dazeley, R. (2012). Recentred local profiles for authorship attribution. *Natural Language Engineering*, *18*(3), 293–312. https://doi.org/10.1017/S1351324911000180

McCarthy, P. M., & Jarvis, S. (2007a). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2007b). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Meara, P. (2005). Lexical frequency profiles: A monte carlo analysis. *Applied Linguistics*, *26*(1), 32–47. https://doi.org/10.1093/applin/amh037

Toolan, M. J. (2009). *Narrative progression in the short story: A corpus stylistic approach*. John Benjamins Pub.

Torruella, J., & Capsada, R. (2013). Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, *95*, 447–454. https://doi.org/10.1016/j.sbspro.2013.10.668

Wimmer, G., & Altmann, G. (1999). Review article: On vocabulary richness. *Journal of Quantitative Linguistics*, *6*(1), 1–9. https://doi.org/10.1076/jqul.6.1.1.4148

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and lexical sophistication in first language writing. In D. Malvern, B. Richards, N. Chipere, & P. Durán (Eds.), Lexical Diversity and Language Development: Quantification and Assessment (pp. 152–176). Palgrave Macmillan UK. https://doi.org/10.1057/9780230511804_9

McCarthy PM. Doctoral dissertation. 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity. Available from Proquest Dissertations and Theses. (UMI No. 3199485)

Emmirich Kelih retrieved on sept 2020 from https://homepage.univie.ac.at/emmerich.kelih/wp-content/uploads/Lexical_richness_Summer_QL_Kelih.pdf