Assessing Assamese Suffix Productivity: A Probabilistic Study in Resource-Limited Contexts

Pinky Moni Gayan[1]
Arup Kumar Nath[2]

Tezpur University

Abstract

Numerous digitally advanced global languages have been studied under the light of morphological productivity; however, Assamese and other Indo-Aryan languages are still understudied in this field, though it is a widely discussed area of morphology. The purpose of this paper is to demonstrate the productivity of 15 suffixes replicated by a few measuring methods in a manually prepared sample. The obtained values are used in the later section to group the suffixes into different clusters based on their similar productivity rate in clustering in R. By determining the general productivity rate of the suffixes from the total productivity rates of all the methods, it demonstrates how clustering in R may be used as an empirical and visual tool for grouping similarly productive suffixes. The paper also reports about the paucity of language resources as well as tools in the language and how bridging this gap could have resulted in more precise, seamless results in a notably shorter amount of time.

Keywords: Morphological productivity; corpus; assamese; cluster in R; natural language understanding, language resources

1. Introduction

A number of prominent scholars have developed theories and ideas, formulated measuring methods and also addressed the issues concerning morphological productivity which have paved the way for further study in this area for upcoming linguists and researchers (Aronoff 1976; Anshen & Aronoff 1989; Baayen 1989, 1992, 1994; Bauer 1992, 2001; Baayen & Lieber 1991; Baayen & Renouf 1996; Plag 1999, 2003 etc.). However, study on measuring morphological productivity based on Indian languages is still rare both in theoretical and descriptive framework. The established quantitative methods of productivity, though helpful, requires digitally-developed corpora. Any study on productivity in any language necessitates large-scale corpora. Corpus-based approach is crucial to understand the nuances of productivity study unlike the dictionary-based approach, hence it is considered as an advanced approach of productivity study (Baayen 1989, 1992, 1994; Bauer 1992, 2001). However, the productivity study in Indian languages lacks well-developed digital corpus like BNC, COCA, etc. The lack of digitally formatted versions of Assamese dictionaries, i.e. *Hemkosh*, *Asamiya Jatiya Abhidhan*, to name a few, hinders the process of a comprehensive study of morphological productivity. The seemingly old approach of dictionary-based study also turns out to be difficult due to the lack of the electronic versions. Therefore, the area of morphology is yet to be explored in many Indian languages.

Through this paper, an attempt is made to explore this phenomenon by extending the usage of established methods of measuring productivity in fifteen Assamese suffixes (Hulse, 2010). It aims to show the outcome of the productivity of the suffixes by measuring methods and their subsequent groupings in Cluster in R. It highlights the lack of full-fledged resources suitable for productivity study, the methodological issues concerning resources, extraction tools and processes.

The selected suffixes for the study are:

-ɔk -অক	*pat^h**ɔk*** ‘reader’, *xɔt**ɔk*** ‘hundred’
-ɔn -অন	*k^{h}aw**ɔn*** ‘the act of eating’, *kɔmp**ɔn*** ‘trembling, shaking’
-ɔna -অনা	*g^{h}ɔt**ɔna*** ‘an accident’, *k^{h}el**ɔna*** ‘toy, playing instrument’
-ɔti -অতি	*bow**ɔti*** ‘flowing’, *nam**ɔti*** ‘one who sings song’
-ɔni- অনী	*row**ɔni*** ‘reaper’, *bow**ɔni*** ‘weaver’
-ɔnija/ -ɔnia -অনীয়া	*poh**ɔnija*** ‘domestic’, *bil**ɔnija*** ‘distributor’
-ɔruwa/*-ɔrua* -অৰুৱা	*batɔr**uwa*** ‘pedestrian’, *hatɔr**uwa*** ‘hat
-al -আল	*mɔŋɔh**al*** ‘fleshy’, *tezal* ‘bloody’
-alu -আলু	*dɔj**alu*** ‘kind’, *krip**alu*** ‘generous’
-aru -আৰু	*zuz**aru*** ‘fighter’, *dub**aru*** ‘diver’
ami -আমি	*t^{h}ɔg**ami*** ‘cheater’, *gor**ami*** ‘orthodox’
-ahi -আহি	*sɔl**ahi*** ‘deceitful’, *mɔd**ahi*** ‘alcoholic’
-ija/-ia -ইয়া	*kumɔl**ija*** ‘not fully grown’, *sɔhɔr**ija*** ‘living in a town’
-ua/-uwa -ওৱা/-উৱা	*g^{h}ɔr**ua*** ‘homely, domestic’, *xar**uwa*** ‘fertile’
-ual/ -uwal -উৱাল	*dak**uwal*** ‘postman’, *pah**uwal*** ‘plumpy’

2. Sample

Studies of morphological productivity predate the easy availability of large-scaled corpora in the target language (Bauer 2001). But when it comes to the study of productivity in Assamese based on the corpora approach, it was an arduous task to find out suitable material or resources that can meet the criteria. First, unlike English, German or any such languages which have seen a growth of advanced digitalisation, a

[1] Department of Linguistics and Language Technology, Research scholar, Tezpur University, Assam. Email: pinkygayan111@gmail.com
[2] Department of Linguistics and Language Technology, Assistant Professor, Tezpur University, Assam. Email: *arupjnu@gmail.com*

regional language like Assamese has a long way to go in this context. Although the language is gradually marking its presence digitally, the resources required for this study are yet to achieve their desired level. While locating pre-processed digital corpus in the language, two digital corpora came into light. One is the EMILLI corpus, which was accessed from the Department of Computer Science and Engineering, Tezpur University (https://www.lancaster.ac.uk/fass/projects/corpus/emille/). EMILLI has been constructed as a part of a collaborative venture between the EMILLI project (Enabling Minority Language Engineering), Lancaster University, UK, and Central Institute of Indian Languages (CIIL), Mysore, India. It consists of three components: monolingual, parallel, and annotated corpora. It contains 14 corpora, including both written and spoken data for 14 South Asian languages: Assamese, Bengali, Gujrati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu, and Urdu. Another corpus named as Assamese general Text Copus was collected from TDIL-DC, CIIL Mysore. However, these two corpora could not be applied to the present study. Although the EMILLI corpus is significantly large and covered texts from different genres, the texts of the corpus are comparatively older which make it unsuitable for the use of studying present productivity of morphological processes. Moreover, in this corpus, Bengali script is used instead of Assamese script in digitization of the texts. Assamese general Text Copus is annotated with POS (Part-of-Speech) tags; however, the design of this corpus did not facilitate the extraction of words corresponding to suffixes. It is therefore evident that although many individuals and institutions are endeavouring to develop corpora in Assamese, still technically full-fledged corpus specifically designed for these kinds of studies is yet to be developed. As a solution, a sample of approximately *one* lakh (0.1 million)[3] words have been collected and compiled for the study. These texts are collected from different online platforms in five different genres- story, article, news, travelogue, and translation. For each section, texts of nearly *twenty thousand* words have been collected.

3. Methodology

Since there was no access to a computational tool in order to identify the words pertaining to the suffixes, the words formed by these suffixes are extracted from the sample by going through the texts manually and also by using the Find Pane computer application available for word documents.

In the process, two primary issues came to the surface. One is because of a characteristic of Assamese orthography and another is inflectional morphemes at the end of a word. Because of the first reason, the above-mentioned search process does not work if we put the particular suffix directly on the 'Search what' box of the 'Search and Replace' dialogue box. For example, লেখক *lek*^{h}*ɔk* 'writer' is formed by attaching -অক -*ɔk* suffix to the base লেখ *lek*^{h}. However, when the suffix is added to the base, the vowel sound অ *ɔ* gets merged with খ *k*^{h}*ɔ* which is the last consonant sound of the base, because of which presence of অ *ɔ* is not visible in orthographic form and if we put -অক -*ɔk* directly in the search box, it does not highlight the words formed by it. To address these orthographic complexities, the most convenient approach for us is to handle it manually by including only the final consonant or the strings of consonants of the particular suffix in the search box, which identifies every single word in the word document that ends with that consonant or the strings of consonants. To identify -অক -*ɔk* suffixed words, -ক *kɔ* was put on the search box while checking the box of 'Match suffix'. It was obvious that a huge number of words gets highlighted in this way and not all the highlighted -ক -*kɔ* ending words are words formed by the suffix -অক -*ɔk*. A few highlighted words such as এক *ek* 'one', নান্দনিক *nandɔnik* *'aesthetic'* etc. are not the result of -অক -*ɔk* suffixation. For this, all the highlighted words are manually checked to eliminate the other non-suffixed words. However, for a more refined extraction process, computational tool developers should collaborate with linguists to create such tools.

Another problem is that the words which end with inflectional suffixes or classifiers do not get highlighted in this method, because of which, the chances of missing a few counts are high. To address this issue, after listing down all the highlighted words formed by this suffix, every word is again searched individually in the Word document to know the exact number of presences of that particular word in the document. For example, if we get *lek*^{h}*ɔk 'writer', gajɔk 'singer', k*^{h}*etijɔk 'farmer'* in the previous search, these words are again being searched individually in the document.

By using the following search strings in the 'Find and Replace' dialogue box, the words are located in the sample for the following *fifteen* suffixes:

1. -অক -*ɔk* : -ক -*kɔ*
2. -অন -*ɔn* : -ন -*nɔ*
3. -অনা -*ɔna* : -না -*na*
4. -অতি/-অনী -*ɔti* : -তি, -তী, -টি, -টী -*ti*, -*ti*, -*ti*, -*ti*
5. - *অনি/- অনী* -*ɔni* : -নি, -নী, -ণি, -ণী -*ni*, -*ni*, -*ni*, -*ni*
6. -অনীয়া/-অনিয়া -*ɔnija*:-নিয়া, -নীয়া, -ণীয়া -*nija*, -*nija*, -*nija*, -*nija*
7. -অৰুৱা -*ɔruwa* : -ৰুৱা, ৰোৱা -*ruwa*, -*rowa*
8. -আল -*al*: -আল -*l*
9. -আৰু -*aru* : -ৰু, -ৰো -*ru*, -*ro*

[3] 100.000 words = A lakh (/læk, lɑːk/; abbreviated L; sometimes written lac) is a unit in the Indian numbering system equal to one hundred thousand (100,000; scientific notation: 10^5).

10. -আমি *-ami* : -মি, -মী *-mi*, *-mi*
11. -ইয়া *-ia* : -য়া *-ja*
12. -উৱা/-ওৱা *-ua* : -ৱা *-wa*
13. -উৱাল, -ওৱাল *-ual* : -ৱাল *-wal*

To measure the productivity of the suffixes, variables of Baayen's measuring methods (Baayen and Lieber 1992; Baayen 1992 etc) are used independently or in combination. These are Type (V) frequency and Token (N) frequency, Type/Token (V/N) or Token/Type (N/V) method, Productivity in the strict sense (p) (n_1/V) and Hapax/Type (V/ n_1) method.

Table 1. Results of Type Frequency, Token Frequency, Hapax Legomena, Token/Type ratio, Type/Token ratio and Token/Type ratio

Sl.	Suffixes	Type (V)	Token (N)	Hapax (n_1)	V/N	N/V	n_1/N	n_1/V
1	-অক *-ɔk*	41	220	17	0.1863	5.365	0.077	0.414
2	-অন *-ɔn*	84	458	37	0.1834	5.452	0.080	0.440
3	-অনা *-ɔna*	28	193	7	0.146	6.892	0.036	0.25
4	-অতি *-ɔti*	9	18	6	0.5	2	0.334	0.667
5	-অনি *-ɔni*	48	159	23	0.301	3.312	0.144	0.479
6	-অনিয়া *-ɔnija*	5	14	1	0.357	2.8	0.071	0.2
7	-অৰুৱা *-ɔruwa*	1	1	1	1	1	1	1
8	-আল *-al*	7	19	3	0.368	2.714	0.157	0.428
9	আলু *-alu*	NIL	-	-	-	-	-	-
10	-আৰু *-aru*	1	2	0	0.5	2	0	0
11	-আমি *-ami*	3	4	2	0.75	1.333	0.5	0.666
12	-আহি *-ahi*	NIL	-	-	-	-	-	-
13	-ইয়া *-ia*	110	418	72	0.263	3.8	0.172	0.654
14	-ওৱা/-উৱা *-ua*	18	51	7	0.352	2.833	0.137	0.388
15	-উৱাল *-ual*	1	2	0	0.5	2	0	0

4. Data and Results

After collecting the data formed by the selected suffixes in the sample, we arrive at the following result (Table 1.) and based on the Table 1., the suffixes are arranged in descending order against each method in Table 2.

Table 2. Ranking of suffixes in descending order of productivity by each method

Ranking	V	N	(n_1)	(V/N) (N/V)	(p) n_1/N	n_1/V
1.	*-ia*	*-ɔn*	*-ia*	*-ɔruwa*	*-ɔruwa*	*-ɔruwa*
2.	*-ɔn*	*-ia*	*-ɔn*	*-ami*	*-ami*	*-ɔti*
3	*-ɔni*	*-ɔk*	*-ɔni*	*-ɔti* *-aru* *-ual*	*-ɔti*	*-ami*
4	*-ɔk*	*-ɔna*	*-ɔk*	*-al*	*-ija*	*-ija*
5	*-ɔna*	*-ɔni*	*-ɔna* *-uwa*	*-ɔnija*	***-al***	*-ɔni*
6	*-uwa*	*-uwa*	*-ɔti*	*-ua*	*-ɔni*	*-ɔn*
7	*-ɔti*	*-al*	*-al*	*-ɔni*	*-ua*	*-al*
8	*-al*	*-ɔti*	*-ami*	*-ija*	*-ɔn*	*-ɔk*
9	*-ɔnija*	*-ɔnija*	*-ɔnija*	*-ɔk*	*-ɔk*	*-ua*

			-ɔruwa			
10	-ami	-ami	-aru-uwal (0)	-ɔn	-ɔnija	-ɔna
11	-ɔruwa -aru	-aru-uwal	-alu-ahi (NIL)	-ɔna	-ɔna	-ɔnija

	-uwal					
12	-alu-ahi (NIL)	-ɔruwa		-alu-ahi (NIL)	-aru-ual (0)	-aru-uwal (0)
13		-alu-ahi (NIL)			-alu-ahi (NIL)	-alu-ahi (NIL)

In the following section, we cluster the suffixes by using clustering in R on the basis of their shared properties. It uses the numeric values of all the measuring methods and based on this, the machine learning algorithm separates them into different clusters. Figure 1, Figure 2 and Figure 3 are the visualisation of the cluster plot in 3, 4 and 5 clusters respectively.

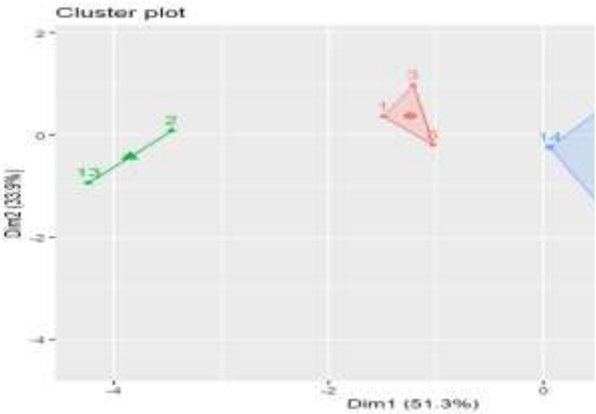

Figure 1: Clustering of the suffixes in 3 clusters by k-means clustering

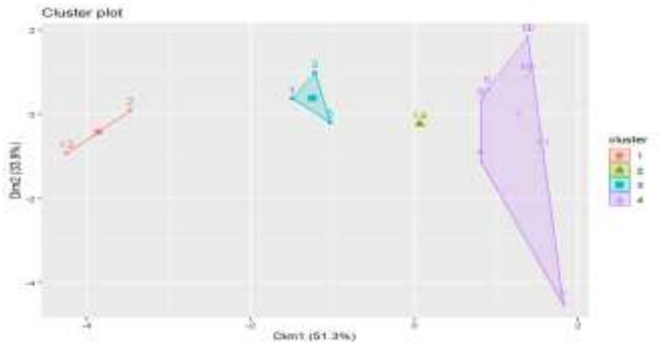

Figure 2: Clustering of the suffixes in 4 clusters by k-means clustering

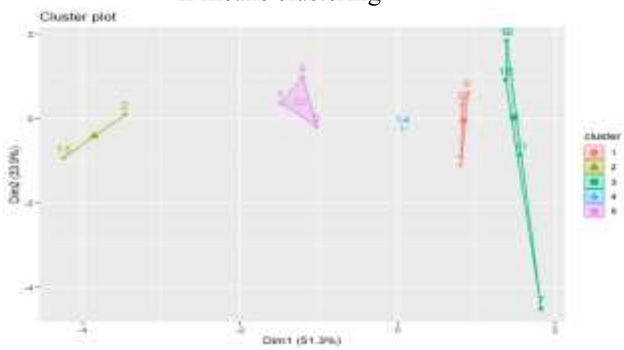

Figure 3: Clustering of the suffixes in 5 clusters by k-means clustering

5. Discussion

The first to note in sample A is that while the results of the first three methods, type (V), token (N), hapax frequency(h1) show similarity and the results of the latter probabilistic measuring methods involving more than one variable is also similar. However, the latter methods display a contrast with the first set of methods.

In the productivity scale, Type frequency V, Token frequency N and hapax frequency, n1 do not show stark differences. Every suffix has altered one step to the higher or lower rank in the methods. When we compare them, we see that *-ia* and *-ɔn* occupy the first and second highest rank in terms of V, whereas *-ɔn* becomes first and *-ia* occupies the second position from the top in N. Similarly, *-ɔni, -ɔk* and *-ɔna* also altered their ranks from 3rd, 4th and 5th in V to 5th, 3rd and 4th in N. *-ɔti* and *-al* occupy 7th and 8th ranking in V and 8th and 7th ranking in N respectively. On the other hand, for the suffixes *-ua, -ɔnija, -ami, -ɔruwa, -aru and -uwal*, the ranking remains unchanged in V and N. From these three, we see that while *-ia* is the most productive suffix in V, *-ɔn* is the most productive one in N. Overall the suffixes *-ia*, *-ɔk, -ɔn, -ɔna and -ɔni* are relatively more productive than the rest in these two methods.

-*ɔruwa*, which has the lowest productivity rate V and N, occupies the highest rank in the probabilistic methods, V/N or N/V, n1/N and n1/V. In table 5.4, the other low-ranking suffixes of V and N, -*ami*, -*aru* and -*uwal* are located higher in the ranking in the probabilistic methods. Similarly, suffixes -*ɔn, -ɔk, -ɔna* which are relatively in the upper position in V and N methods, get a lower rank in the latter. However, a few suffixes -*al, -uwa* relatively remain in the middle position in the productivity scale for all the methods.

From this we came to know that different suffixes are productive in different aspects. Raw counting required for V and N talks about the current and past productivity of suffixes on the basis of existing words, and inferring past and present productivity, the ranking shows that the suffixes -*ia, -ɔn, -ɔk, -ɔni* and -*ɔna* are comparatively more productive than the others in V and N methods.

V/N or N/V, *Productivity in the strict sense (P)* and n_1/V are probabilistic methods and they involve more than one variable in measuring productivity. Hence, they predict the future productivity rate of a suffix, unlike past or present productivity. The suffixes -*ɔti, -ami, -aru, -ɔruwa* appeared more productive only by the latter methods, suggesting that they may have a greater potential for creating new words than the others.

Again, some of the suffixes exhibit somewhat consistent productivity rates across the methods. The productivity of the suffixes -*ia, -ɔni, -ua* and -*ɔti* is slightly higher across the methods, as their distribution can be found from higher to the middle section of the table. These suffixes can be regarded as productive suffixes. The suffix -*ɔk* which is on the upper side in V and N, is found in the lower side in N/V or V/N, and towards the middle position in n1/N and n1/V. As its frequency is on the higher side in all the methods except one, the productivity of this suffix can also be considered as high. Again, the rank of -*al* and -*ɔnija*, on the other hand, can be found in the middle of the hierarchy in all the methods, making them semi-productive suffixes. Nearly in all of the approaches, -*aru, -ual*, -*alu*, -*ahi can* be found in the lower strata, hence indicating them as lowly productive suffixes.

However, one of the classic problems of the probabilistic measuring method is that the extreme number of instances disrupts the true picture. The most productive suffix in the entire sample according to the V/N and N/V, n1/N, and n1/V techniques is -*ɔruwa*, which only has one instance, *dekerua* 'young, full-grown'. The productivity status of a certain suffix is somewhat in doubt if it receives the highest productivity rating while only appearing once in the sample of one lac words. Again, the absence of the two suffixes -*alu* and -*ahi* from sample A does not imply that they are not at all productive in the language. The sample employed here is rather small, and because the suggested statistical approaches are better suited for large-scale corpora, the concerned suffixes do not come across any words in them. This brings up the issue of the lack of well-designed adequate resources in the language, the lack of all-inclusive productivity measurement tools and non-alignment between measuring methods and sampling.

However, regarding the measuring methods, we feel that results by all the measuring methods should be compared to get a comprehensive picture of productivity of the affixes. Also, as different methods display different aspects of productivity, abandoning one method may deprive us of getting other important insights.

Coming to Clustering in R, In Figure 1, the suffixes -*ɔn* and -*ija* are grouped in one cluster, the suffixes -*ɔk*, -*ɔna* and -*ɔni* are grouped in another and the rest of the suffixes clubbed in one when they are clustered in 3.

In Figure 2, where the suffixes are clustered in 4, one more cluster is added for -*uwa*. In Figure 3, in 5 clusters, the next cluster is added by keeping -*ɔti*, -*ɔnija* and -*al* in a separate cluster.

The clustering algorithm of R does the function by forming clusters of variables on their shared properties, i.e., the variables that share similar traits or features are clustered together. Now, we see how the variables which are suffixes, are grouped together and behave similarly. The first observation is that the suffixes of two groups which are formed in the first 3 clustering (Fig 6.5), -*ɔn* and -*ija* in one group and -*ɔk*, -*ɔna* and -*ɔni* in another group are relatively high productive suffixes.

When we change the cluster from 3 to 4, we get -*uwa* in a different cluster, which is also one of the productive suffixes of the language. When the suffixes are separated in 5 clusters, we find another 3 suffixes in one cluster -*ɔti*, -*ɔnija* and -*al* that show a semi-productivity rate. On the other hand, the suffixes that remained in a cluster other than the mentioned above till the 5 clustering, -*ɔruwa, -alu, -aru, -ami, -ahi, -uwal;* we see that they show less productivity rate and they belong to the lower ranking.

Now, the point is that clustering in R accesses the closest similarities of the suffixes based on the numeric values they have in all the methods and groups them accordingly. While we get the productivity rate of the suffixes by using measuring methods and arranging them in a hierarchy for each method separately, clustering in R calculates it by accumulating all the values of all the methods for a suffix and automatically club it with the suffixes which are closely related. That means only the numeric values are considered and the suffixes whose values are found similar were clustered together. Thus, it helps us to prove our claim regarding the productive status of the suffixes. In Figure 3, clusters 2 (-*ɔn*, -*ija*) and 5 (-*ɔk, -ɔna*, -*ɔni*), as we stated, club the suffixes which have similar numeric traits and they are found to be the most frequent suffixes. The same is the case with cluster 4 (-*uwa*) and 1 (*ɔti-, ɔnija, -al*). The third cluster, however, contains the suffixes which have lower frequency rates. We get here that no two suffixes with contrastive values are grouped together, and we can also see that the suffixes from one group have closer ranks in the tables. In terms of productivity, the R clustering presentation is consistent with the labelling of the suffixes of earlier analyses.

6. Concluding Highlights

For a more thorough understanding, the study emphasizes the relevance of Clustering in R for productivity studies. Although the evaluation is not extensive due to the authors having to manually prepare the database, clustering in R is employed as an experimental approach for further, more elaborate studies. Different measuring techniques yield varying productivity rates for the suffixes; in such cases, manual guesswork is necessary to determine the overall productivity tendency of the suffixes. In this context, clustering in R accomplishes the same goal and works as intended empirically. It permits the suffixes to be grouped together according to the characteristics they have in common. We observe that similar clusters are typically formed by suffixes with similar frequencies. The methodology section in Section 3 of the study provides further details on the data extraction process, thereby highlighting the necessity of modifying or improving the Assamese resources that are currently available. As we can see, the methodological process not only takes a considerable amount of time in the absence of the necessary resources, but it also leaves room for several flaws in the data extraction process. It highlights the necessity of creating corpora in Indo-Aryan languages or a segmentation tool in a way that facilitates or permits these kinds of studies in the future.

7. Acknowledgements

We thank Tezpur University for giving us this opportunity to conduct this research work; the informants for providing data and their valuable time; the faculties of the department of Linguistics and Language Technology, Tezpur University for their endless support, kind and insightful comments; we thank each and everyone who have directly or indirectly extended their helping hands to us. We thank Tezpur University administration, University Grant Commission, India and Ministry of Social Justice and Empowerment, Govt. of India for granting fellowship that has provided financial support to this work. We also offer our thanks to Tonmoya Bhattacharya, Department of Computer Science and Engineering, Tezpur University for her kind assistance in writing the R-script.

8. References

Anshen, F. & M. Aronoff (1989). Morphological productivity, word frequency and the Oxford English Dictionary. In R.W. Fasold & D. Schiffrin (Eds.). pp197- 202.

Aronoff, M. (1976). Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass*, (1), 1-134.

Baayen, R. H. (1989). *A Corpus-based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistics Interpretation.* Dissertation. Vrijei University, Amsterdam.

Baayen, R.H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. Van Marle (Eds.),*Yearboook of morphology*, pp 109-149 .Dordrecht: Kluwer

Baayen, R.H. (1994). Productivity in Language Production. *Language and Cognitive Processes*, Vol.9 (3): pp 447-469.

Bauer, L. (1992). Scalar productivity and –lily adverbs. In G. Booij & J. Van Marle (Eds.), *Yearboook of morphology*, pp185-191. Dordrecht: Kluwer

Bauer, L. (2001). *Morphological Productivity*. Cambridge: Cambridge University Press.

Baayen, R. H. and R. Lieber. (1991). 'Productivity and English Derivation: A Corpus Based Study'. *Linguistics* 29, 801–843.

Baayen, R.H. & A. Renouf (1996). Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language*, Vol.72 (1): pp 69-96.

Plag, I (1999). *Morphological productivity: structural constraints in English derivation*. Topics in English Linguistics. Berlin; New York: Mouton de Gruyter.

---------- (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.

Hulse, V. (2010). *Productivity in Morphological Negation: a corpus-based approach*. The University of Manchester (United Kingdom). Retrieved from https://www.research.manchester.ac.uk/portal/files/54505284/FULL_TEXT.PDF . Accession date 30/08/2021