CHAPTER 2

Research Framework and Methodology

Chapter 2 deals with the description of the methodologies and methods used in this investigation. It describes the methodological strategy that is used to collect, organize, and evaluate the data in order to accomplish the research objectives and answer the research questions.

2.1 Research approach and design

The study follows a descriptive design to interpret the phenomenon of productivity. It utilizes a hybrid method, which integrates both qualitative and quantitative approaches. It begins by gathering, measuring, and analysing data for the affixes. Even if we may infer intuitively that some affixes are more productive than others; or based on semantic or structural aspects that may indicate their productivity, we still need to provide empirical evidence to demonstrate their actual use in everyday language. This objective can be achieved and data inference can be aided by statistical measurements. To develop a more thorough grasp of a study subject, it aims to gather insights and comprehend how the affixes act as well as the 'why' behind the statistical data.

2.2 Data Collection: Sample Size

In the first chapter, it is mentioned that for the study, corpus-based and dictionary-based approaches are adopted. The first one, the corpus-based approach of study, is one of the major approaches to study morphological productivity. Although the studies of morphological productivity require large-scale corpus, when it comes to the study of productivity in Assamese based on the corpora approach, it was an arduous task to find suitable material or resources that can meet the criteria. First, unlike English which has seen a growth of advanced digitalization, a 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 purpose are yet to achieve their desired level. Regarding the corpus data, it has been tried to locate pre-processed digital corpus in the language, and in this process, we came to know about two digital corpora. One is the EMILI corpus, which was accessed from the Department of Computer Science and Engineering, Tezpur University. (https://www.lancaster.ac.uk/fass/projects/corpus/emille/). EMILI has been constructed

as a part of a collaborative venture between the EMILI 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 fourteen corpora, including both written and spoken data for fourteen South Asian languages: Assamese, Bengali, Gujrati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Shimla, Tamil, Telugu, and Urdu. Another digital corpus of written language was collected from CIIL Mysore, India. Both corpora, however, proved to be unsuitable for this use. Even though the EMILI corpus was considerably large and included texts from a variety of genres that satisfied the requirements for corpus size, the majority of the texts are relatively older and do not belong to the contemporary era, making the corpus of little use for researching the productivity of morphological processes today. Moreover, in this corpus, Bengali script was used instead of Assamese script in the digitization of the texts. Another corpus was annotated with POS (Part-of-Speech) tags, which also failed to fulfil the requirement. It is soon found that although many individuals and institutions are endeavouring to develop corpora in Assamese, still an appropriate corpus specifically designed for these kinds of studies is yet to be developed.

As a solution, two samples are created for the study, the first sample, which is sample A, is created by collecting and compiling data from digital texts consisting of one lac words for the study. These texts are collected from different online platforms in five different genres- story, article, news, travelogue, and translation. For each section, nearly twenty thousand words have been collected. It is already established that the study of productivity requires a large-scale corpus and the bigger a corpus is, the more convenient the result would be. But the lack of such a full-fledged corpus in the language bound us to compile samples on our own.

Another sample, which is sample B, is prepared from a prominent dictionary of Assamese *Hemkosh*. The dictionary sample comprises the data from two editions of Hemkosh 2006 and 2016.

2.3 Data collection procedure

2.3.1 Sample A: Corpus

Coming to sample A, the texts are collected from digital platforms and compiled in a single word document. We lacked access to a computational tool in order to identify the words based on their affixes. All the words by these affixes are extracted from the sample simply by going through the texts manually and by using the Find Pane computer application available for word documents.

After this, to locate all the different words formed by affixes and their total numbers, the following steps are followed:

- Type ‘Ctrl + Find’
- Click on ‘Advanced Find’

The ‘Find and Replace’ dialogue box appears here.

- Type the strings of letters on ‘Find what’
- Click on ‘More’ to expand the dialogue box
- Click on ‘Match suffix’/ Click on ‘Match prefix’
- Go to ‘Reading Highlight’
- Click on ‘Highlight all’

It then displays the number of total words containing those strings of letters and highlight those words in the word document. After that, the highlighted words are manually scanned. However, not every word that ends with a sound similar to a certain suffix or begins with a sound similar to a prefix is actually a suffixed or prefixed word. For example, in *ɔnustʰan* ‘function’, *ɔ-* is not a prefix here, rather it is only a string of the word. That is why, all the highlighted words are manually checked to eliminate the other non-prefixed and non-suffixed words.

During the process of gathering words for suffixes, two problems emerged, which include one being a feature of Assamese orthography and the other being inflectional morphemes at word ends. In Assamese orthography, because a consonant sound

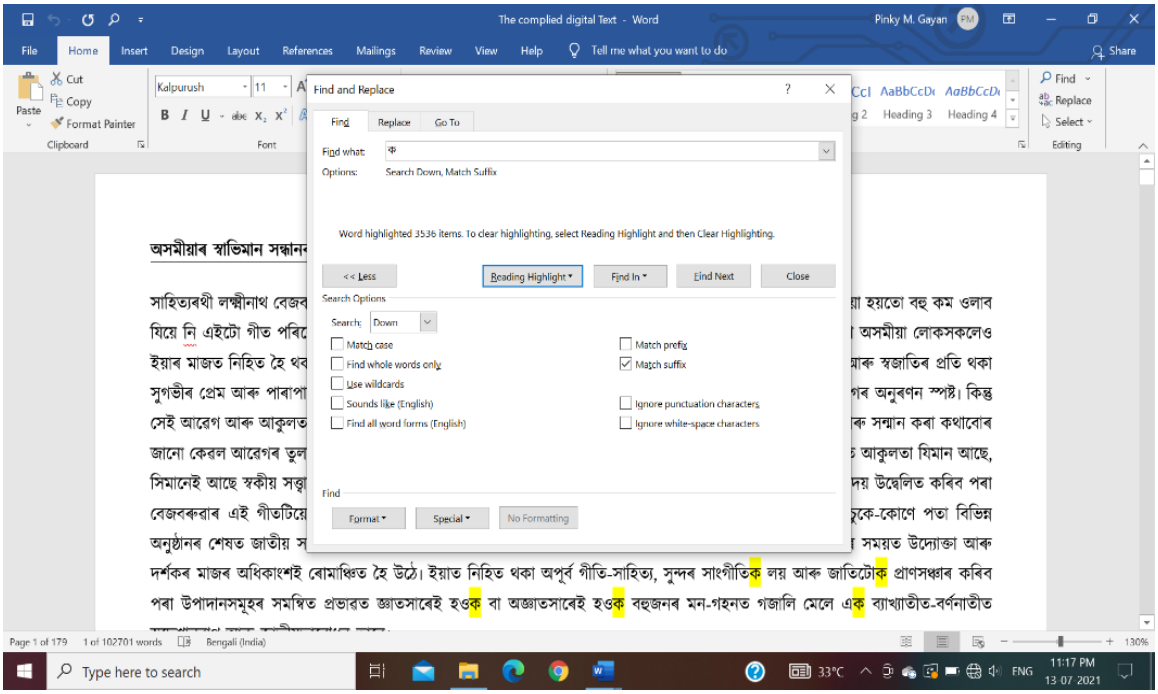
inherently bears the vowel sound, the vowel sound in a suffix is realised alongside the root or base; as a result, in orthographic form, the vowel sound in a suffix is not isolated from the root or base. Because of this, the intended result cannot be obtained by merely typing the suffix in the 'Search what' dialogue box. For example, while searching for the words created by the suffix -*ɔk* -অক, we were unable to highlight the words formed by the same as -*ɔ* is present in conjunction with the last consonant sound of the root or base. As an instance, *lekʰɔk* লেখক 'writer' is formed by attaching -*ɔk* অক suffix to the base *lekʰ* লেখ. But if we directly put -অক -*ɔk* in the search box, it is unable to highlight the word লেখক *lekʰɔk* in the document. To overcome this issue, only the last consonant or the strings of consonants of that suffix are given in the search box, which identifies every single word in the word document which ends with that consonant or the strings of consonants. To identify -*ɔk* -অক suffixed words, *kɔ* ক is put on the search box while checking the box of 'Match suffix'. It is obvious that a huge number of words gets highlighted in this way and not all the highlighted *kɔ* ক ending words are words formed by the suffix -অক. For example, the highlighted *ek* এক 'one', *nandɔnik* নান্দনিক 'beautiful' etc. words are not -*ɔk* -অক suffixed words. For this, all the highlighted words are manually checked to eliminate the non-suffixed words.

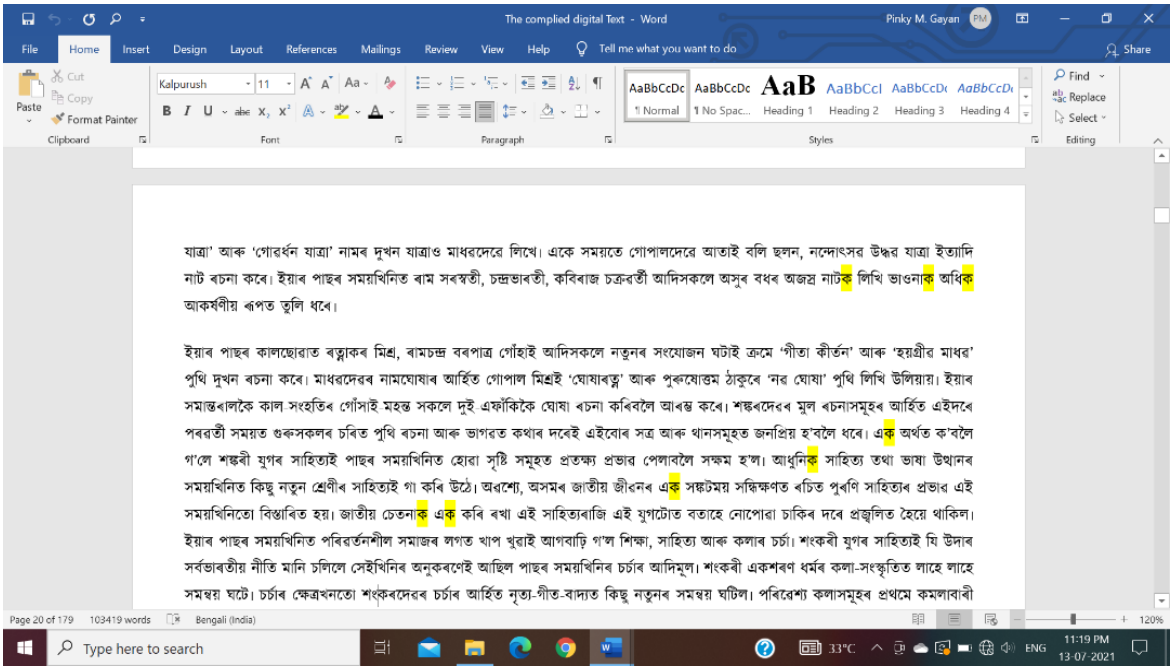
A further problem with the data extraction process is that words containing a chosen suffix that end in inflections or classifiers are not highlighted, increasing the likelihood of missing a few counts. In order to solve this problem, all the highlighted words created by the affixes are listed, and then each word is searched in the Word document once more to determine how many times it appears there. For example, if we find *ɔdʰɔrmɔ* 'misdeed, sin' and *pohɔnija* 'domestic' in the previous search, then those words are again searched in the document separately to determine the number of tokens.

For the investigation, we have chosen fifteen suffixes and six prefixes. While prefixes can be used directly in search strings to locate prefixed words, to highlight the words formed by the suffixes, the following search strings are used against each suffix:

1. -অক *-ɔk* : -ক *-kɔ*
2. -অন *-ɔn* : -ন *-nɔ*
3. -অনা *-ɔna* : -না *-na*
4. -অতি *-ɔti* : -তি, -তী, -টি, -টী *-ti*
5. -অনি *-ɔni* : -নি, -নী, -ণি, -ণী *-ni*
6. -অনিয়া *-ɔnija* : -নিয়া, - নীয়া, -ণিয়া, -ণীয়া *-nija*
7. -অৰুৱা *-ɔruwa* : -ৰুৱা, -ৰূৱা, -ৰোৱা *-rowa*
8. -আল *-al* : -ল *-lɔ*
9. -আলু *-alu* : -লু *-lu*
10. -আৰু *-aru* : -ৰু, -ৰূ, -ৰো *-ru/ -ro*
11. -আমি *-ami* : -মি, -মী *-mi*
12. -আহি *-ahi* : -হি, -হী *-hi*
13. -ইয়া *-ija/-ia* : -য়া *-ja*
14. -ওৱা/-উৱা *-ua/-uwal* : -ৱা *-wa*
15. -উৱাল *-ual/-uwal* : -ৱাল *-uwal*

The following two images are attached to demonstrate how the process is carried out:





2.3.2 Sample B: Dictionary

For sample B, which consists of the data collected from the dictionary, we have taken two editions of *Hemkosh*, ed. 2006 and ed. 2016. Next, starting with the 2016 version of the dictionary, all the words for each affix are manually counted to prepare the data set. Then, using the same procedure, we also counted the words in the 2006 editions. While counting the words of the 2006 edition, we have marked the words in the previous dataset that are added to the most recent edition but are not included in the 2006 edition of *Hemkosh*. In this way, we are able to determine how many new words have been added over the course of ten years.

2.3.3 Selection and Elimination Criteria

The kind of words that are appropriate for quantitative productivity measurement or that can truly be considered as suitable members for determining the nature of productivity have not been determined yet; therefore, a few criteria based on which data are selected and eliminated are stated here. Before being used for study, data must be carefully collected and evaluated because inaccurate data, such as misspelled words, repeated articles, and sections, affect frequency distributions. (Evert and Ludeling 2001).

As this study concerns derivational affixes, it follows some criteria to select and eliminate words from the respective resources.

We mainly found three different sorts of bases with derived affixes that can be categorized as roots or non-roots. Bases with other existing affixes and compound bases fall within the non-root category. While compiling the data set, the words where an affix is attached to a root or base with other preexisting affixes are the ones that are gathered first. Compounds and other forms are excluded primarily due to the possibility that the productivity of affixes in compound words or in any other word-formation processes may produce a different result which can itself be another area to be studied under the light of productivity measurement.

The criteria are listed below:

i. To count a word, the base should be independent, a bound stem, (*a non-independent word),* a derived base containing other existing derivational affixes. Compound words are excluded from the listing.

Affixes attached to independent bases:

অধর্ম *ɔdʰɔrmɔ* = অ + ধর্ম *ɔ* 'prefix' + *dʰɔrmɔ*
পাঠক *patʰɔk* 'a reader' = পাঠ + অক *patʰ* 'lesson' + *ɔk* 'suffix'
বুজন্ *buzɔn* 'understanding' = বুজ + অন *buz* 'to understand' + *ɔk* 'suffix'
সুদর্শন *xudɔrxɔn* 'handsome' = সু + দর্শন *xu* 'prefix' + *dɔrxɔn* 'sight'
কুমলীয়া *kumɔlija* 'not fully developed'= কোমল + -ঈয়া *komɔl* 'soft' + *ija* 'suffix'

ii. Affixes attached to bases with preexisting affixes:

প্রদর্শক *prɔdɔrxɔk* 'An exhibitor' = প্র + দৃশ্ + **-অক** *prɔ* 'prefix' + *drix* 'to see' + *ɔk* 'suffix'
পৰিব্রাজক *pɔribrazɔk* 'A tourist' = পৰি + ব্রজ + **-অক** *pɔri* 'prefix' + *brɔz* 'to travel' + *ɔk* 'suffix'
বিচক্ষণ *bisɔkʰjɔn* 'skillful' = **বি**- + চক্ষ্ + **-অন** *bi* 'prefix' + *sɔikʰ* 'to talk'+ *ɔn* 'suffix'

iii. Words that accidentally start or end with the same letters or word forms are eliminated.

Example: অকণ *ɔkɔn* ‘A little, a few’

iv. Proper names are excluded.

Example: দুর্যোধন *durjodʰɔn*

The dictionary contains a few words whose bases are dependent on their affixes to function independently. They are in the process of lexicalization or fossilization in the language. Although it was initially decided that these words would not be counted, this presents some questions. It is also from these lexicalized forms that speakers derive ideas about the usefulness of the affixes. It should be noted that the identical affixes that are affixed to these non-independent bases are also accessible with independent bases. In such scenario, it is the non-independent bases, rather than the relevant affixes that we presume have lost productivity in isolation. However, the following considerations have been made regarding the inclusion of words having non-independent bases:

- The preference will be given to the words, where an affix and base can easily be segmented.
- If the base is not a commonly identifiable or usable word, firstly, the separate existence of this will be checked in the dictionary. If it is present, then it is counted.
- If the meaning of such bases is cited in the dictionary in segmented form, even if its separate existence is not listed in the dictionary, then also it will be included.

Example: *ɔkɔtɔla* ‘The state of not removing the branches and leaves’ = *ɔ* ‘prefix’ + *k̃ɔtal* ‘to curb’. Here, although the base *k̃ɔtal* is not given a separate entry in the dictionary, the meaning of it is mentioned alongside *ɔkɔtɔla*. Therefore, this word is counted in the dataset. But, if the presence and meaning of the non-independent base cannot be traced anywhere, it has been decided to abandon those words. It is because, in this case, it is concluded that the word is lexicalized or fossilized fully and the affix no longer serves as an affix in that word.

However, although it is stated that compounds are excluded from the calculation while doing so, a few considerations have been made depending on the nature of the compounds. The excluded compounds are:

a) The hyphenated compound words are excluded, as in this case, the words can stand alone most of the time.

Example: *dati-kaxɔrija = dati + kax + ɔr + ija*
Dwelling in a frontier = Edge + near + suffix + suffix
ɔtitʰi-xewɔk = ɔtitʰi + xewa + ɔk
Guest servants = guests + service + suffix

b) Quantitative and directive compounds are excluded.

Example: *dudinija = du + din + ija*
Of two days = two + day + suffix
x̃opʰɔlija = xo + pʰal + ija
Of right side = right + side + suffix

However, the dataset contains a certain group of compounds, which comprise lexicalized compounds that have been assimilated and are now an integral part of native speakers' vocabulary. Since the affixes are now an essential component of it, speakers are no longer aware of or use unaffixed forms.

Example: *kɔlpɔtua* 'sheath of a banana plant' = *kɔl* 'banana' + *pat* 'leaf' + *ua*

pɔtʰalisɔkua 'Having horizontal eyes (said of human beings)' = *pɔtʰali* 'horizontal' + *sɔku* 'eyes' + *ua*.

2.4 Issues/limitations of data collection

The process of collecting and arranging data for the study of morphological productivity raises a number of issues. The fundamental methodological challenge is the lack of pre-processed, well-developed corpora, as productivity studies generally call for the availability of sizable databases or corpora. Although dictionary-based, corpora-based, and psycholinguistic testing are all distinct approaches, they are occasionally

combined for comparison purposes. In this study, two established approaches are decided to exploit to address the issue of productivity in Assamese, both of which required manual sampling.

The segmentation of affixes is a problem with the corpora technique. As far as we are aware, one of the main difficulties in NLP is the lack of an auto-segmentation tool that can segment the morphemes or at the very least can recognize the affixes after they have undergone morphophonemic modifications following attestation. The other challenges are brought by homophonous morphemes, meaning ambiguity, semantic non-transparency, and the absence of etymology information, that make it impossible to categorize the morphemes distinctly.

However, a language must be improved digitally in order to exist in this day and age. This initial phase of this study demonstrates technological limitations of the language. Even if a language has an abundance of existing literary resources, it must now compete in the digital sphere in order to survive. This ends up being the largest obstacle or limitation for productivity research.

Due to the unavailability of digital dictionaries and corpora, a significant amount of manual effort must be used to gather and organize the data. It takes a long time to do this as well. Indian languages are falling behind in this area of morphology research because of this issue. Avoiding a problem does not always result in a solution, thus it must be addressed in order to investigate it further. A large database or corpus is required for the study to produce an accurate result as prior research in this field has demonstrated that size of the corpus matters. For the study of production in Assamese, this presents the biggest obstacle. It is crucial to note that despite the demand for the issue, we must place some limitations on ourselves because it is not humanly possible to develop a very large corpus and collect data from numerous dictionaries manually. The database built for the dictionary approach and the corpus would only be viewed as representative samples of the study, not as exhaustive.

2.5 Analysis techniques: The measuring methods

Notion of productivity has always been an issue of debate amongst linguists and hence the endeavour is still on to find out the best measuring methods which can predict

the productivity of a word-formation process maximally. When the concept of measuring productivity emerged, the need for statistical tools also came to surface. A lot of methods have been formulated or introduced in the literature by different scholars for measuring productivity. However, to date, an all-encompassing method is still missing. It is because the concept of productivity is still multidimensional. Again, each measure establishes a different ranking of productivity, and the proposed measuring methods predict different aspects of productivity (Plag 2003). Baayen (1993) finds that affixes may rank quite differently depending on which measures are used. Hence, based on the approach different methods need to be employed. Not only that, each method also has different drawbacks and methodological issues of data sampling and analysis.

The two samples we have taken for the study are entirely different as sample A is a collection of contemporary texts and sample B is the dataset prepared from a dictionary, hence applicable methods are not entirely the same. For the corpus-based approach, it has been decided to employ the variables of Baayen's measuring methods independently or in combination to examine the nature of productivity (Hulse, 2010). These are: Type frequency and Token frequency, Type/Token or Token/Type method, Productivity in the strict sense p* (Hapax/Token) and Hapax/Type method. However, for the dictionary-based approach, we could find only one applicable method, that is, the Type frequency method.

2.5.1 Type Frequency (V)

The simplest method of measuring productivity is Type and Token Frequencies (Baayen and Lieber 1991). Type is a distinct symbol, which means it refers to a type of 'symbol' such as 'A' or 'B' or 'C'. *Type frequency* means the counting of different types of words or lexemes in a corpus. Every new word is considered as a new type and the accumulation of all the distinct types occurring in a sample or corpus is called Type Frequency (Baayen and Lieber 1991; Plag et el 1999; Bauer 2001; Plag 2003, 2006). The bigger the number of types for an affix, the higher the productivity of that affix is. Type frequency focuses on recording the number of different words that are produced or coined by using a particular process. In that way, it helps to identify the maximum number of words that are created within a period. However, it can only provide factual

productivity, which means it cannot predict the rate at which an affix can be used in new word formation in the future. It can only talk about the past productivity of an affix. Bolozky (1999) advocates type-based methods for measuring productivity, especially for dictionary-based approaches.

2.5.2 Token Frequency (N) method

Token frequency (N) means the number of all the occurrences by a type or by an affix in a text or corpus, be it in different words or in repeated occurrences of the same word. (Aronoff 1976, 1980; Anshen and Aronoff 1988; Baayen and Lieber 1991). Like Type frequency, in Token frequency also the higher token frequency indicates higher productivity. Token frequency says how commonly or frequently the types are used in real-world situations.

2.5.3 Type/Token (V/N) or Token/Type (N/V) method

While in the Type/Token method, the higher ratio means higher productivity, in the Token/Type method, a higher ratio indicates lower productivity (Baayen and Lieber 1991; Plag 1999, Hulse 2010). The reason behind this is assumed that as productive affixes are frequently involved in new word formations, not all the words are used in a wide manner all the time because new words are constantly entering into the lexicon (Baayen and Lieber 1991).

However, type frequency and token frequency (2.5.1 and 2.5.2) individually do not say much about the possibility of forming new words in the future. Type and token frequency are like predictors and cannot straightforwardly be related to productivity (Bauer 2001). This problem led the scholars to formulate probabilistic methods which involve more than one variable in measuring productivity. Baayen and his collaborators (Baayen 1992, 1993, Baayen and Lieber 1991, Baayen and Renouf 1996, and elsewhere) have developed a range of measuring methods, the central predictor or variable of which is hapax legomena along with types and tokens.

2.5.4 Productivity in the strict sense (P) or n1/n method

Hapax legomena or hapax means the type of word that occurs only once in the entire text or corpus. Often Hapax is viewed as an indicator of productivity (Baayen and Lieber 1991; Baayen and Renouf 1996, Baayen, 1992, Plag 2003), because hapaxes are the lowest frequency types that occur in a corpus where we can find most of the neologisms (Cited in Plag 2003). Although Hapaxes cannot be declared as neologisms directly; because of their lowest frequency, there is a high probability that most of the words of this category are newly formed words by a word-formation process. Again, the size of the corpora also plays a crucial role in determining the nature of hapaxes. If the size of the corpora is small, then the most hapax legomena are known words of the language and if the corpus is large, it is seen that most of them are neologisms (Plag 2003). However, in this research work, the authors adopt a small sample of manually collected texts with one lakh words only, therefore, the hapax or the words which occur only once in the entire sample is an established word of the language.

The measuring method *Productivity in the strict sense (P)* (Baayen 1992, Baayen 1993, Baayen and Lieber 1991) also known as *category-conditioned degree of productivity* utilizes the Hapax legomena to predict the productivity of a morphological process. It is also known as 'potential productivity', as it shows the approximate growth rate of vocabulary. According to this measure, the higher frequency is associated with a higher degree of productivity.

The formula for calculating productivity by this method is $(P) = n_1/N$

Here, P = productivity, n1= number of hapaxes by a suffix and N= Token frequency of that particular suffix.

2.5.5 Hapax/Type (n_1/V) method

This method utilizes hapax legomena and the number of types, which is calculated by following the formula n1/V. In this method also, a higher productivity is signified by the higher Hapax/Type ratio. It is correlated with type frequency to measure productivity. Van Marle (1992) argues that Type frequency is more useful than Token frequency to

gauge productivity, as Type frequency allows or records the words that are produced by utilising a morphological process. Similar to Marle, Bybee (2003) also opines that type rather than token frequency underlies productivity.

2.6 Clustering in R

After measuring the suffixes by using statistical methods, a machine algorithm called Clustering in R is employed to cluster the suffixes. R is a versatile open-source programming language for statistics and data science. It is a scripting language for statistical data manipulation and analysis (Matloff 2011). Machine learning algorithms are typically divided into groups according to the kind of output variable and the kind of issue that needs to be solved. These algorithms can be broadly categorized into three types: classification, clustering, and regression. While clustering is a sort of unsupervised algorithm, regression and classification are examples of supervised learning algorithms.

Following a quantitative examination of the productivity rate of the affixes using a few measuring techniques, they are grouped together to evaluate if there is a relationship among them; in this case, the clustering algorithm of R serves the purpose. The purpose of incorporating clustering in R for the productivity study is to verify the validity of the productivity status and determine whether the output of R aligns with the results obtained from statistical methods, semantic analysis, and other structural properties.

Clustering is an unsupervised machine-learning algorithm. It is used to create clusters from data points with comparable features. The data points in a cluster should ideally have similar characteristics, while the points in separate clusters should have as distinct characteristics as feasible. In this work, R clustering clusters or groups the affixes based on similar properties, these properties are presented in terms of numerical values calculated by the previous statistical methods.

There are several types of clustering methods including K-means clustering, Hierarchical clustering, DBSCAN clustering, BRICH, spectral clustering etc. However, among these, K-means clustering method is chosen for grouping the affixes, as it is one of the primary and relatively simple methods. Moreover, clustering of the affixes is

included as a supporting measure to substantiate our conclusions, rather than as a component of a thorough investigation.

The *k-means* clustering contains a data set, which means a collection of data values in a matrix format. In other words, it is a collection of values in columns and rows in the form of variables and observation respectively. A 'variable' is all the measured values of the same underlying attribute, i.e., an attribute or characteristic of the observation. The value of each measuring method across the affixes, this way, is a variable in the study. On the other hand, an observation is a case of the collected data or all values that were measured for the same unit. The value of each affix across the measuring methods of our study is an observation.

Another feature of clustering in R is that the number of clusters needs to be provided beforehand in which the variables would be grouped. However, the ideal number of ideal clusters is considered as three (Matloff 2011). Nonetheless, the number of clusters might be raised to verify alignment with the outcome.