The Median Threshold Hypothesis : Measuring morphological productivity from frequency lists

Gauvain Schalchli
Université Bordeaux-Montaigne/CLLE

1 Context: A decline of frequency based studies on quantitative productivity because of methodology process limitations

Morphological productivity can be studied from a theorical point of view with attention focused on constraints on rules/processes/schemas application. However, another important point of view is the quantitative study of the extent of use of the morphological units. A first approach of quantitative studies is based on the type frequency of the morphological units. Specifically, this approach focused on new types in diachrony (Aronoff and Lindsay 2014; Berg 2020), neology (Cartier et al. 2018) or contemporary synchrony (Dal and Namer 2012, 2015; Dal et al. 2018). A different approach, inspired from corpus linguistics and psycholinguistics, is based on token frequency. In that approach, two aspects of productivity of morphological categories are captured: 1) the extent of new formations (the constitutive aspect) 2) lexicalized idiosyncratic items (the limitative aspect). Token frequency is estimated from the number of occurrences of the lexical units in a large and representative corpus.

The frequency-based quantitative approach of morphological productivity has been developed during the 90's from the works of Harald Baayen (Harald Baayen 1989, 1991, 1992a, 1992b, 1993, 1994, 1996, 2001, 2002; Harald Baayen and Lieber 1991; Chitashvili and Baayen 1993; Harald Baayen and Renouf 1996; Harald Baayen and Neijt 1997; Harald Baayen and Tweedie 1998; Plag, Dalton-Puffer, and Baayen 1999; Hay and Baayen 2002, 2003). He principally developed an index measure named potential productivity and defined, for one morphological process, the ratio between hapax number of the process and its cumulative frequency of occurrence (Baayen 2009; Gaeta and Ricca 2015; Dal and Namer 2016). This measure has been applied on a large scale for English (Baayen and Lieber 1991), Italian (Gaeta and Ricca 2003, 2006) and Dutch (Baayen 1989).

In French, the use of potential productivity index began in the first decade of the twenty-first century (Dal 2003). During that period, it has been applied on different special cases, like *-et/-ette* suffixation (Fradin, Hathout, and Meunier 2003), suffixes *-ité* and *-able* comparison (Grabar et al. 2006), comparison between *-able*, *-ité* et *-is(er)* suffixes (Namer 2003), denominal adjectival suffixes (Grabar and Zweigenbaum 2003), compound nouns (Voskovskaia 2009). However, the major study has concerned only 8 different processes (Dal et al. 2008).

At the same time of the first French studies, strong evidence has been produced of subtantial limitations of the Baayen's index. (Evert and Lüdeling 2001) shows that calculations based on automatic procedures are very different of those based on manual procedure. (Hay 2001) shows that relative frequency of bases impacts productivity. (Gaeta and Ricca 2003a, 2003b, 2006) show that: 1) the interaction between prefixation and suffixation in derivation cycles plays a role in productivity calculations; 2) productivity comparison between different processes with Baayen's index is only available for equal cumulative frequency.

Overall, these observations clarified and reinforced potential productivity, but also limited its scope and made its procedure more cumbersome. In correlation with that effect, we can

observe an important slowdown in potential productivity studies. In Harald Baayen research, the last innovative paper on potential productivity dates from the beginning of the first decade of twenty-first century (Hay and Baayen 2003).[1] In French, the large scale project coordinated by Georgette Dal and initiated by the "morphological productivity" team of the GDR 2220 has been abandoned and the majority of works on productivity after 2008 deals with other quantitative strategies (Dal and Namer 2010; Koehl 2010, 2012; Cartier et al. 2018; Missud, Amsili, and Villoing 2020)[2]. We can observe the same tendency at the international level. From the two more recent reviews on productivity (Dal and Namer 2016; Gaeta and Ricca 2015), the last original works on potential productivity cited (Dal et al. 2008; Gaeta 2007) date from the first decade of twenty-first century (excepted some applied works like (Chmielik and Grabar 2011; Vendrell and Domínguez 2012; Wieling et al. 2014). In our own bibliographical exploration, we just found just three unreferenced works on potential productivity (Hennecke and Baayen 2017; Voskovskaia 2013, 2019). Moreover, the majority of research on productivity attempts to develop alternative strategies (Fernández-Domínguez 2010; Säily 2011, 2016; Berg 2020). The last but not least indication of that slowdown, the last international handbook on morphology (Audring and Masini 2019) contains no specific chapter on productivity, and only briefly cites potential productivity in the sub-chapter 3.5.2 "Productivity and blocking" (Lieber 2019).

2 Methodology: Facilitating the estimation of productivity from frequency lists by the Median Threshold Hypothesis

The fundamental intuition about frequency-based estimation of productivity is that the high part and the low part of the scale don't represent the same aspect of morphological knowledge (Fernández-Domínguez 2010). It is explicitly argued by Baayen (1992:110): "Any measure of morphological productivity [...] will have to satisfy a number of requirements. [...] such a measure should express "the statistically determinable readiness with which an element enters into new combinations." [...] taking into account these formations which are characterized by formally or semantically idiosyncratic properties should have the effect of lowering the value of the productivity measure." We propose to call the high frequency part of a lexicon his head and the low frequency part his tail.

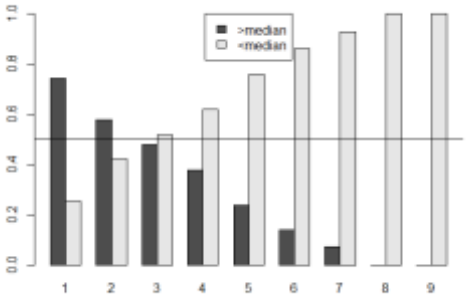
Because it is inspired from biological probabilistic models (Chitashvili and Baayen 1993), Baayen chose the hapax count as an estimator for neologisms, however, not all hapaxes are neologisms and not all neologisms are hapaxes. On the other hand, the choice of cumulative frequency as estimator of high frequency items is another approximation. For example, (Baayen, Wurm, and Aycock 2007) use a threshold of 6 occurrences per million rather than the unique hapax rank for defined probable neologisms. Potential productivity estimate productivity with the minimal part of the lexicon tail and an extensive but confusing estimator of the lexicon head.

From a descriptive (VS inferential) statistical view on lexical frequency data, information about productivity extracted from high frequency items VS low frequency items could embrace the entire frequency scale. In order to avoid all groundless theoretical hypotheses, we propose to cut the lexicon and its frequency scale in two equal parts centered on the

[1] Further works of Harald Baayen did'nt answer the Gaeta & Ricca's discussion and did'nt apply the variable corpus approach (e.g. (Denistia and Baayen 2019; Shen and Baayen 2021))

[2] See also Dal & Namer (2012), Dal et al (2018)

median rank. From this point of departure, the estimation of productivity consists of comparing the number of instances of a morphological process in the head of the scale (over the median) with those present in the tail (under the median). The higher the number of the instances of the process in the tail, the more likely that the process is productive. Likewise, the higher the number of instances of the process in the head, the less likely that its productivity is strong.
In the following, we use the lexical frequency list of Lexique3 extracted from a 50 million words corpus of film subtitles and whose occurence counts are strongly correlated with lexical decision times (New et al. 2007). This list contains approximately 40,000 lexemes with phonological transcriptions. For that selection, the median rank frequency is 0.39 occurrences per million. 148 entries have exactly that frequency, 19,501 have a higher frequency than the median: this the head of the lexicon, while 20,382 have a lower frequency than the median: this is the tail of the lexicon. If we count the number of word forms more and less frequent than the median for each number of syllables by word, we find the barplot below:



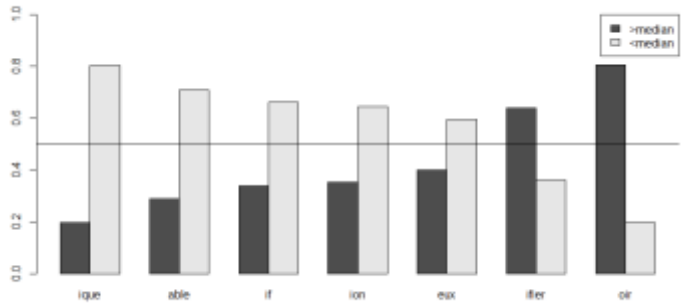
This barplot shows that short words belong mainly in the head and long words mainly in the tail. This observation deals with productivity, as it is well known that constructed words are typically longer than unconstructed words. We can express the relationship between the two sub-populations of a class of lexemes (e.g. one-syllable lexemes) by a ratio of T/H, similar to that of potential productivity, where T is the tail population and H is the head population. Applied on morphological categories, we can interpret the ratio in term of productivity. If the ratio is around 1, productivity level is medium. The more the ratio increases from one, the more the process is productive, and likewise unproductive in the contrary case. For example, in our sample, the ratios for one-syllable and two-syllable words which are rarely a result of productive morphological processes are approximately 0.3 and 0.7 whereas that of three-syllables and four-syllables syllable words which are more frequently morphologically constructed are approximately 1.2 and 1.75.

3 Results: Testing the hypothesis by discriminating French suffixal word endings from non-suffixal word endings

On the one hand, the median threshold hypothesis functions as a null hypothesis in statistical tests and allows us to differentiate between productive and non-productive processes or lexical properties: if the ratio of the numbers of instances of a process on either side of the median (T/H) is close to zero, then the index postulates that the process is not productive. If, on the contrary, the ratio is significantly greater than 0, then it must be postulated that the process is productive to some degree.
Moreover, the productivity index based on the median threshold hypothesis can give rise to different cases that logically split the productivity spectrum by acting as an index of the degree

of productivity. If the tail of the vocabulary contains no instantiations of the process under study, the index is equal to 0, which is equivalent to null productivity. If the tail of the vocabulary contains as many instances of the process as the head, then the productivity is equal to 1, which corresponds to a significant productivity. Between these two first values, we can interpret the index in a gradual way: the closer the index is to zero, the lower the productivity and the closer it is to 1, the more significant it is. Finally, if the number of instances in the tail is higher than the number of instances in the head, then the productivity is high and the further away from 1 the higher it is.

Dal et al (2008) classify seven suffixes in three levels of productivity based on potential productivity index. From their counting, *-able* and *-ique* are highly productive, *-eux*, *-if*, *-ion* and *-ifier* are moderately productive and *-oir* has a low level of productivity. Applying our index to the data of Lexique3[3] shows a comparable classification of these seven suffixes:



-ique and *-able* have the higher tail proportion of attestations and *-oir* have the higher head proportion. About the ratios, *-ique* and *-able* are comparable to five-syllables words. *-if*, *-ion* and *-eux's* ratios are comparable to four-syllables, *-ifier* to two-syllable words and *-oir* to one-syllable words.

In order to validate the hypothesis on a large scale, we will present its application to different lexical categories and to different suffixes and morphological problems like allomorphy and competition from french data and from different corpora.

5 Conclusion

We propose a new quantitative estimation of productivity, comparable to Baayen's potential productivity but avoiding its shortcomings. The T/H ratio is based on frequency of occurrence, it is easy to compute from a list of frequencies, it is robust against corpus size variation and against automatic morphological analysis, it allows the comparability of all processes whatever their frequency of occurrence in the corpus, it takes into account in a consistent and interpretable way the effect of high frequencies on lexicalization and the relevance of low frequencies for morphology.

However, this proposal is simplistic and many refinements may be considered in order to advance the modeling of morphological productivity and to adapt it to different contexts or varieties. First of all, other thresholds can be easily tested. Second of all, it is consistent with

[3] For this example, we worked without manual validation of the morphological analysability and without category selection.

diachronic estimates and surveys from occasionalisms and non-conventional corpora such as the web. Finally, it allows us to imagine a dynamic interpretation of productivity as a function of the saturation phase of the derivational domain of the measured construction.

References

Aronoff, Mark, and Mark Lindsay. 2014. 'Productivity, Blocking, and Lexicalization'. *The Oxford Handbook of Derivational Morphology*. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199641642.001.0001/oxfordhb-9780199641642-e-005 (April 16, 2021).

Audring, and Masini, eds. 2019. The Oxford Handbook of Morphological Theory *The Oxford Handbook of Morphological Theory*. Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199668984.001.0001/oxfordhb-9780199668984 (April 17, 2021).

Baayen, H., Lee H. Wurm, and Joanna Aycock. 2007. 'Lexical Dynamics for Low-Frequency Complex Words: A Regression Study across Tasks and Modalities'. *The Mental Lexicon* 2(3): 419–63.

Baayen, Harald. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation*.

———. 1991. 'A Stochastic Process for Word Frequency Distributions'. In *29th Annual Meeting of the Association for Computational Linguistics*, , 271–78.

———. 1992a. 'Quantitative Aspects of Morphological Productivity'. In *Yearbook of Morphology 1991*, Springer, 109–49.

———. 1992b. 'Statistical Models for Word Frequency Distributions: A Linguistic Evaluation'. *Computers and the Humanities* 26(5): 347–63.

———. 1993. 'On Frequency, Transparency and Productivity'. In *Yearbook of Morphology 1992*, Springer, 181–208.

———. 1994. 'Productivity in Language Production'. *Language and Cognitive Processes* 9(3): 447–69.

———. 1996. 'The Effects of Lexical Specialization on the Growth Curve of the Vocabulary'. *Computational Linguistics* 22(4): 455–80.

———. 2001. 'Word Frequency Distributions'. *Text, speech and language technology; 18*.

———. 2002. 'Affix Ordering and Productivity: A Blend of Phonotactics and Prosody,

Frequency, and Lexical Strata'. In *Yearbook of Morphology 2001*, Yearbook of Morphology, eds. Geert Booij and Jaap Van Marle. Dordrecht: Springer Netherlands, 181–82. https://doi.org/10.1007/978-94-017-3726-5_6 (March 30, 2021).

———. 2009. '43. Corpus Linguistics in Morphology: Morphological Productivity'. *Corpus linguistics. An international handbook*: 900–919.

Baayen, Harald, and Rochelle Lieber. 1991. 'Productivity and English Derivation: A Corpus-Based Study'. *Linguistics* 29(5): 801–44.

Baayen, Harald, and Anneke Neijt. 1997. 'Productivity in Context: A Case Study of a Dutch Suffix'.

Baayen, Harald, and Antoinette Renouf. 1996. 'Chronicling the Times: Productive Lexical Innovations in an English Newspaper'. *Language*: 69–96.

Baayen, Harald, and Fiona J. Tweedie. 1998. 'Sample-Size Invariance of LNRE Model Parameters: Problems and Opportunities'. *Journal of Quantitative Linguistics* 5(3): 145–54.

Berg, Kristian. 2020. 'Changes in the Productivity of Word-Formation Patterns: Some Methodological Remarks'. *Linguistics* 58(4): 1117–50.

Cartier, Emmanuel et al. 2018. 'Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain'. *SHS Web of Conferences* 46: 08002.

Chitashvili, R. J, and R. H. Baayen. 1993. 'Word Frequency Distributions'. In *Quantitative TextAnalysis*, Wissenschaftlicher Verlag Trier, eds. G. Altmann and L. Hrebicek. , 54–135.

Chmielik, Jolanta, and Natalia Grabar. 2011. 'Détection de La Spécialisation Scientifique et Technique Des Documents Biomédicaux Grâce Aux Informations Morphologiques'. *TAL* 51(2): 151–79.

Dal, Georgette. 2003. *La Productivité En Questions et En Expérimentations*. Larousse.

———. 2008. 'Quelques Préalables Au Calcul de La Productivité Des Règles Constructionnelles et Premiers Résultats.' In *Congrès Mondial de Linguistique Française*, EDP Sciences, 142.

———. 2018. 'Toile versus dictionnaires : Les nominalisations du français en-age et en-ment'. *SHS Web of Conferences* 46: 08003.

Dal, Georgette, and Fiammetta Namer. 2010. 'Les Noms En-Ance/-Ence Du

Français: Quel (s) Patron (s) Constructionnel (s)?' *2ème Congrès Mondial de Linguistique Française*: 060.

———. 2012. 'Faut-il brûler les dictionnaires ? Ou comment les ressources numériques ont révolutionné les recherches en morphologie'. *SHS Web of Conferences* 1: 1261–76.

———. 2015. '133. Internet'. https://halshs.archives-ouvertes.fr/halshs-02275998 (April 16, 2021).

———. 2016. 'Productivity'. In *The Cambridge Handbook of Morphology*, ed. Andrew Hippisley & Gregory T. Stump. Cambridge University Press., 70–90. https://hal.archives-ouvertes.fr/hal-01303313 (December 18, 2020).

Denistia, Karlina, and R. Harald Baayen. 2019. 'The Indonesian Prefixes PE- and PEN-: A Study in Productivity and Allomorphy'. *Morphology* 3(29): 385–407.

Evert, Stefan, and Anke Lüdeling. 2001. 'Measuring Morphological Productivity: Is Automatic Preprocessing Sufficient'. In *Proceedings of the Corpus Linguistics 2001 Conference*, UCREL, 167–75.

Fernández-Domínguez, Jesús. 2010. 'Productivity vs. Lexicalization: Frequency-Based Hypotheses on Word-Formation'. *Poznań Studies in Contemporary Linguistics* 46(2): 193–219.

Fradin, Bernard, Nabil Hathout, and Fanny Meunier. 2003. 'La Suffixation En-ET et La Question de La Productivité'. *Langue française*: 56–78.

Gaeta, Livio. 2007. 'On the Double Nature of Productivity in Inflectional Morphology'. *Morphology* 17(2): 181–205.

Gaeta, Livio, and Davide Ricca. 2003a. 'Frequency and Productivity in Italian Derivation: A Comparison between Corpus-Based and Lexicographical Data'.

———. 2003b. 'Italian Prefixes and Productivity: A Quantitative Approach'. *Acta Linguistica Hungarica* 50(1–2): 93–112.

———. 2006. 'Productivity in Italian Word Formation: A Variable-Corpus Approach'.

———. 2015. 'Productivity'.

Grabar, Natalia et al. 2006. 'Productivité Quantitative Des Suffixations Par-Ité et-Able Dans Un Corpus Journalistique Moderne'. In *TALN*, , 167–75.

Grabar, Natalia, and Pierre Zweigenbaum. 2003. 'Productivité à Travers Domaines et Genres: Dérivés

Adjectivaux et Langue Médicale'. *Langue française*: 102–25.

Hay, Jennifer. 2001. 'Lexical Frequency in Morphology: Is Everything Relative?' *Linguistics*.

Hay, Jennifer, and Harald Baayen. 2002. 'Parsing and Productivity'. In *Yearbook of Morphology 2001*, Springer, 203–35.

———. 2003. 'Phonotactics, Parsing and Productivity'. *Italian Journal of Linguistics* 15: 99–130.

Hennecke, Inga, and Harald Baayen. 2017. 'A Quantitative Survey of N Prep N Constructions in Romance Languages and Prepositional Variability'. *Quaderns de filología. Estudis lingüístics* (22): 129–46.

Koehl, Aurore. 2010. 'Les Noms de Propriété Adjectivale En-Eur et-Esse: Un Modèle Évolutif Original'. *2ème Congrès Mondial de Linguistique Française*: 066.

———. 2012. 'La Construction Morphologique Des Noms Désadjectivaux Suffixés En Français'. PhD Thesis. Université de Lorraine.

Lieber, Rochelle. 2019. 'Theoretical Issues in Word Formation'. *The Oxford Handbook of Morphological Theory*. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199668984.001.0001/oxfordhb-9780199668984-e-3 (April 17, 2021).

Missud, Alice, Pascal Amsili, and Florence Villoing. 2020. 'VerNom: Une Base de Paires Morphologiques Acquise Sur Très Gros Corpus (VerNom: A French Derivational Database Acquired on a Massive Corpus)'. In *Actes de La 6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition), Traitement Automatique Des Langues Naturelles (TALN, 27e Édition), Rencontre Des Étudiants Chercheurs En Informatique Pour Le Traitement Automatique Des Langues (RÉCITAL, 22e Édition). Volume 2: Traitement Automatique Des Langues Naturelles*, , 305–13.

Namer, Fiammetta. 2003. 'Productivité Morphologique, Représentativité et Complexité de La Base: Le Système MoQuête'. *Langue française*: 79–101.

New, Boris, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. 'The Use of Film Subtitles to Estimate Word Frequencies'. *APPLIED PSYCHOLINGUISTICS* 28(4): 661–77.

Plag, Ingo, Christiane Dalton-Puffer, and Harald Baayen. 1999. 'Morphological Productivity across Speech and Writing'. *English Language & Linguistics* 3(2): 209–28.

Säily, Tanja. 2011. ‘Variation in Morphological Productivity in the BNC: Sociolinguistic and Methodological Considerations’. 7(1): 119–41.

———. 2016. ‘Sociolinguistic Variation in Morphological Productivity in Eighteenth-Century English’. *Corpus Linguistics and Linguistic Theory* 12(1): 129–51.

Shen, Tian, and R. Harald Baayen. 2021. ‘Adjective–Noun Compounds in Mandarin: A Study on Productivity’. *Corpus Linguistics and Linguistic Theory*. https://www.degruyter.com/document/doi/10.1515/cllt-2020-0059/html (June 25, 2021).

Vendrell, Mercedes Roldán, and Jesús Fernández Domínguez. 2012. ‘Emergent Neologisms and Lexical Gaps in Specialised Languages’. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 18(1): 9–26.

Voskovskaia, Elena. 2009. ‘Morphological Productivity and Family Size: Evidence from French Compound Nouns Garde-x and N-de-N’. In *Mediterranean Morphology Meetings*, , 123–33.

———. 2013. ‘La Productivité Des Noms Composés En Français Du XVIIe Au Début Du XXe Siècle’. PhD Thesis.

———. 2019. ‘Composés NN et NA Dans La Littérature Française Du 17e Au 20e Siècle: La Productivité Morphologique’. In Paris.: Université Paris Diderot.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R. Harald Baayen. 2014. *Lexical Differences Between Tuscan Dialects and Standard Italian: A Sociolinguistic Analysis Using Generalized Additive Mixed Modeling.*