Published in final edited form as:
Stud Health Technol Inform. 2019 August 21; 264: 1433–1434. doi:10.3233/SHTI190470.

P-Hacking Lexical Richness Through Definitions of “Type” and “Token”

K. Bretonnel Cohen[a], **Lawrence E. Hunter**[a], **Peter S. Pressman**[b]

[a]Computational Bioscience Program, University of Colorado School of Medicine, Aurora, Colorado, USA

[b]Department of Neurology, University of Colorado Hospitals, Anschutz Medical Campus, Aurora, Colorado, USA

Abstract

“P-hacking” is the repeated analysis of data until a statistically significant result is achieved. We show that p-hacking can also occur during data generation, sometimes unintentionally. We use the type-token ratio to demonstrate that differences in the definitions of “type” and “token” can produce significantly different results. Since these terms are rarely defined in the biomedical literature, the result is an inability to meaningfully interpret the body of literature that makes use of this measure.

Keywords

Language; vocabulary

Introduction

There is a growing awareness that many published scientific results are not reproducible. Statistical practices are one of the most common factors blamed for this “reproducibility crisis,” as it has come to be known [1]. There are many aspects of the contributions of statistical malpractice to the reproducibility crisis, including an increasingly recognized practice known as p-hacking. *P-hacking* is the act of repeatedly re-analyzing data until a statistically significant result is achieved [1; 2]. One method of p-hacking is varying aspects of the hypothesis tests themselves until a positive finding is reached; another is reselecting subsets of the data. However, there are other ways that p-hacking can be done, without “cherry-picking” hypothesis tests *or* data – as this paper shows, it can also be done at the data generation phase. This kind of p-hacking is especially pernicious for at least four reasons: (1) it leaves no trace in the form of questionable statistical assumptions that might otherwise be noticed by a reviewer, (2) it leaves no trace even when using publicly available data sets, since the data itself is unchanged, (3) it is not difficult to justify to oneself, and (4) publication practices make it quite easy. The approach in question is especially relevant to the computational bioscience community because it is easily illustrated with a measure that is widely used in neuroscience, psychiatry, developmental psychology, and neurology: the type-token ratio.

We illustrate the issues with the concept of "lexical richness" because its calculation seems very straightforward but is quite nuanced in practice. Informally, *lexical richness* refers to the quality, variability, and sophistication of the vocabulary of a speaker or a text. It is usually defined by some form of the *type-token ratio*, calculated as the number of distinct words (*types*) divided by the total number of words (*tokens*). It is commonly used in several biomedical fields [3]. Despite the prevalence of this measure in the scientific literature, it is unclear how to interpret the literature using the type-token ratio because papers typically do not define what they are counting as types and tokens, and there is no consensus definition for either.

Publication bias in the literature on diagnostic use of measures of lexical richness may promote p-hacking and unreliable results. We know that the tendency is not to be able to publish negative results [4]. Do people sometimes modify those "simple" decisions perfectly innocently, without actually realizing their linguistic consequences, until they get a positive finding on a subsequent statistical hypothesis test—a positive finding that would in fact be a negative finding if they made different preprocessing decisions? It is not difficult to do so, because "type" and "token" are generally not defined when used in the literature or elsewhere. We present modeling results that explore the implications of failures to consistently define those variables. If we find that modifying the definitions of "type" and "token" can affect whether or not statistically significant differences are found in the type-token ratio, then it should be clear that these terms *must* be defined when using them. If, in fact, they are *not* typically defined, then there is a problem in our field, and we should be aware of it and take action by being explicit about those definitions.

Materials and Methods

The materials are drawn from the CRAFT corpus of biomedical journal articles [5]. We modeled two different definitions of *type* and three different definitions of *token*. Definitions of "type" unavoidably interact with definitions of "token", which can result in different classifications of the same term depending on precedence orderings, so we processed all data sources using 10 random orderings of random subsets of five sets of type and token definitions. We used the conservative (nonparametric) Wilcoxon signed rank test to look for statistically significant differences between the distributions of type-token ratios in the outputs. The null hypothesis is that there is no effect of different definitions of the variables. For our definitions, we used observations from three comprehensive analyses of tokenization [6–8]. See the code available on GitHub for how we handled phrasal verbs (e.g. *have_to* versus *have to*), negative clitics (e.g. *do n't* versus *don't*), and repetitions. Type definitions applied here include normalized or unnormalized digit types or letter cases, punctuation inclusion in or exclusion from the token counts, and different treatments of hyphens and underscores. (See [9] for how contentious any of these decisions can be.) There are interactions between any definition of type or token, so we ran different combinations of the various definitions of each and their orderings, all randomized, for 10 combinations of definitions and orderings. Type-token ratio was calculated as the count of types divided by the count of tokens for the complete text.

Analysis of the resulting distributions follows the approach of [10] (see also [8; 11]). Figure 1 shows the distributions of the type-token ratios for 10 randomly selected permutations of type definition, token definition, and order of application. Results cluster into groups with two widely differing magnitudes, one just below a type-token ratio of 0.5 (1, 3, 5, and 10) and the other around 0.2 (2, 4, 6, 7, 8, and 9); clearly, the definitions of type and token can have an enormous effect on the magnitude of the type-token ratio. In no case are the type-token ratios normally distributed, meaning the type-token ratio is vulnerable to Type I errors from parametric hypothesis tests.

The distributions of the permutations with the highest and lowest median type-token ratios in the low-magnitude group were significantly different ($p < 0.001$); within the high-magnitude group, however, they were not. We also tested all pairs of permutations within the two groups. Table 1 shows that more than 40% of the definition set pairs did not yield statistically significant differences, while nearly 60% did, suggesting that there are many ways that statistically significant differences can be found (or not) based only on the different definitions of “type” and of “token”. Like typical p-hackers, we did no multiple testing correction.

The complete code base and outputs of all steps of the analysis are available on GitHub.

Discussion

These results demonstrate that there are many “minor” preprocessing decisions that can be the difference between having statistically significant results and not, independent of any actual differences in the underlying data.

Conclusions

The precise definitions of “type” and “token” are so important that, if they are absent, research findings based on the type-token ratio are essentially uninterpretable. In fact, precise definitions of these terms are rarely given in the biomedical literature — even in work on tokenization [11]! We have shown that type-token ratios can be significantly different based only on differences in how “type” and “token” are defined. These results suggest that the relevant literature should be read with caution. Talking about his initial work on this problem. J. Simmons, the originator of the term “p-hacking,” says that “we realized entire literatures could be false positives” [12]. If research using measures of lexical richness is to avoid turning out that way, future work in the field should always give the precise definitions of “type” and “token” that were used.

References

[1]. Head ML, Holman L, Lanfear R, Kahn AT, and Jennions MD, The Extent and Consequences of P-Hacking in Science, PLOS Biology 13 (2015), e1002106.

[2]. Bishop DVM and Thompson PA, Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value, PeerJ 4 (2016), e1715. [PubMed: 26925335]

[3]. Fergadiotis G, Wright HH, and West TM, Measuring lexical diversity in narrative discourse of people with aphasia, Am J Speech Lang Pathol 22 (2013), S397–408. [PubMed: 23695912]

[4]. Wilson C, Kerr D, Noel-Storr A, and Quinn TJ, Associations with publication and assessing publication bias in dementia diagnostic test accuracy studies, Int J Geriatr Psychiatry 30 (2015), 1250–1256. [PubMed: 25779466]

[5]. Cohen KB, Verspoor K, Fort K, Funk C, Bada M, Palmer M, and Hunter L, The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation In The Biomedical Domain, in: Handbook of Linguistic Annotation, 2016.

[6]. Tomanek K, Wermter J, and Hahn U, A reappraisal of sentence and token splitting for life sciences documents, Stud Health Technol Inform 129 (2007), 524–528. [PubMed: 17911772]

[7]. Trieschnigg D, Kraaij W, and de Jong F, The influence of basic tokenization on biomedical document retrieval, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 803–804.

[8]. Cruz Diaz NP and Maña López M, An Analysis of Biomedical Tokenization: Problems and Strategies, in: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 40–49.

[9]. Née É, Méthodes et outils informatiques pour l’analyse des discours, Presses universitaires de Rennes, 2017.

[10]. Habert B, Adda G, Adda-Decker M, de Marëuil PB, Ferrari S, Ferret O, Illouz G, and Paroubek P, Towards tokenization evaluation, in: Proceedings of LREC, 1998, pp. 427–431.

[11]. He Y and Kayaalp M, A Comparison of 13 Tokenizers on MEDLINE, The Lister Hill National Center for Biomedical Communications 48 (2006).

[12]. Dominus S, When the Revolution Came for Amy Cuddy, in: The New York Times Magazine, New York Times, New York, 2017.

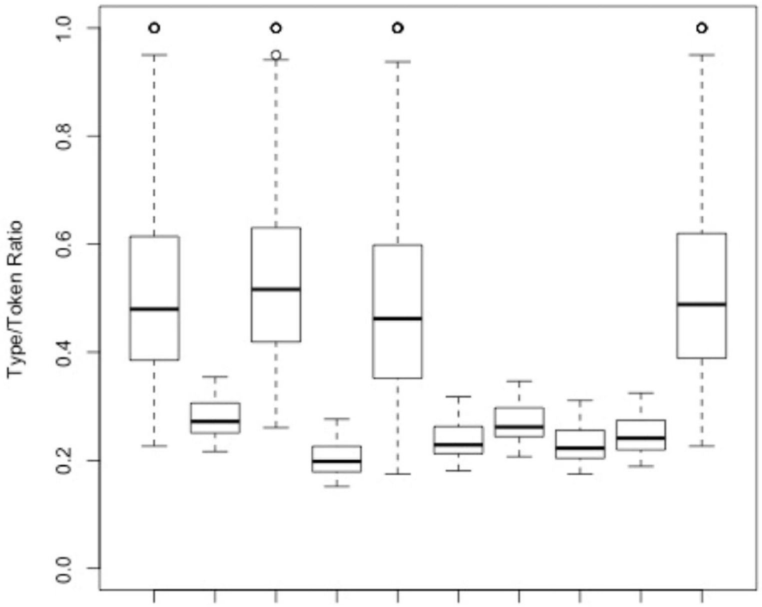

Figure 1 –.

10 random permutations of subsets of the definitions of type and token. Each data point is the type-token ratio for one paper. Each column represents one permutation.

Table 1 –

Number of pairwise differences between sets of definitions of type and token, divided into groups with high- and low-magnitude type-token ratios.

Group	Pairs	Significantly different (%)
High	7	0 (0%)
Low	16	13 (81%)
All pairs	23	13 (57%)