Review History

RSOS-200008.R0 (Original submission)

Review form: Reviewer 1

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept with minor revision (please list in comments)
Comments to the Author(s)
In this work the authors study statistical regularities in natural language. Specifically, they consider samples from written text and study the growth of the vocabulary (Heaps' law). Using analytical calculations and computational analysis of 75 different books, they find substantial differences between Heaps curves of different word classes (nouns, verbs, ...).

The latter aspect constitutes a new contribution to the analysis of Heaps' law in the context of statistical laws in natural language. The findings are substantiated by the statistical analysis and the employed methodology is sound. While the authors cannot provide an explanation for their (admittedly curious) findings, the empirical findings alone will serve as a starting point for future analysis. The manuscript is written very clearly and the careful selection of figures make it easy to follow the different steps in the analysis.

Therefore, I fully recommend publication of the current manuscript.

I would only ask for minor revision in terms of the description of the methodology in order to ensure the reproducibility of the analysis. First, the description of the pre-processing of the data seems insufficient. For example, when using the NLTK-tokenizer, did the authors filter any words? Second, I couldnt find any information on the repository where the code for the analysis of the data will be published (criteria for publication state that "Datasets, code, and other digital materials should be deposited in an appropriate, recognised, publicly available repository").

Review form: Reviewer 2

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept with minor revision (please list in comments)

Comments to the Author(s)
This manuscript investigates Heaps' law in literary texts. The main contribution of this manuscript is the analysis of the words classified by different parts of speech. As far as I am aware this is the first manuscript that performs this analysis, which adds a meaningful contribution to this area of study. The manuscript is clearly written and the statistical analysis, including the comparison to null models, are correctly performed. I recommend the manuscript for publication after the authors address the points listed below.

1) I found the second sentence of the abstract unclear, and the whole abstract over complicated. I suggest trying to simplify this sentence and to focus on the main results.
2) It would be helpful if the authors would use the distinction between word types and word tokens, which is standard in linguistics. For instance, in the caption of Fig. 2 it is sometimes hard to distinguish which concept is being referred (Also in other parts of the manuscript).

3) The main result of the manuscript in Fig. 2 is very interesting. For large values of $N_{\text{tag}}$ the newly added $V_{\text{tag}}$ must correspond to very rare words, possibly including words that are not in standard dictionaries or list of words. These results rely heavily on the POS tagger and it'd be in general important to add more information about how it is tagging the words. It is remarkable that the POS tagger seem to consistently tag these words as nouns, verbs, and others. It'd be helpful if the authors would add some information about how the POS tagger works and whether it can be trusted even for extremely rare words? Should the scaling be expected even for $N \rightarrow \infty$ and is the tagging reliable in this limit?

4) In Sec. 4 and in the third paragraph of the discussion 6 the authors discuss the comparison to randomized texts. This is an important part of the manuscript, which is indeed very relevant and contains original contribution. However, I believe that the observation that randomized texts show larger $V_{\text{tags}}$ is not new. It has been observed and derived mathematically in Ref. 22, but probably it was known even earlier. It is natural to expect taking into account that the usage of words in the text is correlated, with words clustering in regions of the text. This correlation delays the appearance of new words. It is not present in shuffled text, which therefore have a larger number of distinct word tokens for texts of similar length. I believe the authors should revise their claims of novelty in this aspect, specially in the Discussion.

5) All data is available and the Dryad server is working. However, it is not clear how to replicate the results because there is no code available for the data filtering or for the NLP part. The books in the repository still contain metadata, such as translator's preface, and it is not clear how the authors performed the filtering (if any).

**Decision letter (RSOS-200008.R0)**

12-Feb-2020

Dear Dr Zanette,

On behalf of the Editors, I am pleased to inform you that your Manuscript RSOS-200008 entitled "Heaps' law and Heaps functions in tagged texts: Evidences of their linguistic relevance" has been accepted for publication in Royal Society Open Science subject to minor revision in accordance with the referee suggestions. Please find the referees' comments at the end of this email.

The reviewers and handling editors have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the comments and revise your manuscript.

- **Ethics statement**
  If your study uses humans or animals please include details of the ethical approval received, including the name of the committee that granted approval. For human studies please also detail whether informed consent was obtained. For field studies on animals please include details of all permissions, licences and/or approvals granted to carry out the fieldwork.

- **Data accessibility**
  It is a condition of publication that all supporting data are made available either as supplementary information or preferably in a suitable permanent repository. The data
accessibility section should state where the article's supporting data can be accessed. This section should also include details, where possible, of where to access other relevant research materials such as statistical tools, protocols, software etc can be accessed. If the data has been deposited in an external repository this section should list the database, accession number and link to the DOI for all data from the article that has been made publicly available. Data sets that have been deposited in an external repository and have a DOI should also be appropriately cited in the manuscript and included in the reference list.

If you wish to submit your supporting data or code to Dryad (http://datadryad.org/), or modify your current submission to dryad, please use the following link: http://datadryad.org/submit?journalID=RSOS&manu=RSOS-200008

• Competing interests
Please declare any financial or non-financial competing interests, or state that you have no competing interests.

• Authors’ contributions
All submissions, other than those with a single author, must include an Authors’ Contributions section which individually lists the specific contribution of each author. The list of Authors should meet all of the following criteria; 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published.

All contributors who do not meet all of these criteria should be included in the acknowledgements.

We suggest the following format:
AB carried out the molecular lab work, participated in data analysis, carried out sequence alignments, participated in the design of the study and drafted the manuscript; CD carried out the statistical analyses; EF collected field data; GH conceived of the study, designed the study, coordinated the study and helped draft the manuscript. All authors gave final approval for publication.

• Acknowledgements
Please acknowledge anyone who contributed to the study but did not meet the authorship criteria.

• Funding statement
Please list the source of funding for each author.

Please ensure you have prepared your revision in accordance with the guidance at https://royalsociety.org/journals/authors/author-guidelines/ -- please note that we cannot publish your manuscript without the end statements. We have included a screenshot example of the end statements for reference. If you feel that a given heading is not relevant to your paper, please nevertheless include the heading and explicitly state that it is not relevant to your work.

Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript before 21-Feb-2020. Please note that the revision deadline will expire at 00.00am on this date. If you do not think you will be able to meet this date please let me know immediately.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions". Under "Actions," click on "Create a Revision." You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.
When submitting your revised manuscript, you will be able to respond to the comments made by the referees and upload a file "Response to Referees" in "Section 6 - File Upload". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referees. We strongly recommend uploading two versions of your revised manuscript:

1) Identifying all the changes that have been made (for instance, in coloured highlight, in bold text, or tracked changes);
2) A 'clean' version of the new manuscript that incorporates the changes made, but does not highlight them.

When uploading your revised files please make sure that you have:

1) A text file of the manuscript (tex, txt, rtf, docx or doc), references, tables (including captions) and figure captions. Do not upload a PDF as your "Main Document";
2) A separate electronic file of each figure (EPS or print-quality PDF preferred (either format should be produced directly from original creation package), or original software format);
3) Included a 100 word media summary of your paper when requested at submission. Please ensure you have entered correct contact details (email, institution and telephone) in your user account;
4) Included the raw data to support the claims made in your paper. You can either include your data as electronic supplementary material or upload to a repository and include the relevant doi within your manuscript. Make sure it is clear in your data accessibility statement how the data can be accessed;
5) All supplementary materials accompanying an accepted article will be treated as in their final form. Note that the Royal Society will neither edit nor typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details where possible (authors, article title, journal name).

Supplementary files will be published alongside the paper on the journal website and posted on the online figshare repository (https://rs.figshare.com/). The heading and legend provided for each supplementary file during the submission process will be used to create the figshare page, so please ensure these are accurate and informative so that your files can be found in searches. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Please note that Royal Society Open Science charge article processing charges for all new submissions that are accepted for publication. Charges will also apply to papers transferred to Royal Society Open Science from other Royal Society Publishing journals, as well as papers submitted as part of our collaboration with the Royal Society of Chemistry (https://royalsocietypublishing.org/rsos/chemistry).

If your manuscript is newly submitted and subsequently accepted for publication, you will be asked to pay the article processing charge, unless you request a waiver and this is approved by Royal Society Publishing. You can find out more about the charges at https://royalsocietypublishing.org/rsos/charges. Should you have any queries, please contact openscience@royalsociety.org.

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Kind regards,
Lianne Parkhouse
Editorial Coordinator
Reviewer: 1

Comments to the Author(s)

In this work the authors study statistical regularities in natural language. Specifically, they consider samples from written text and study the growth of the vocabulary (Heaps' law). Using analytical calculations and computational analysis of 75 different books, they find substantial differences between Heaps curves of different word classes (nouns, verbs, ...).

The latter aspect constitutes a new contribution to the analysis of Heaps' law in the context of statistical laws in natural language. The findings are substantiated by the statistical analysis and the employed methodology is sound. While the authors cannot provide an explanation for their (admittedly curious) findings, the empirical findings alone will serve as a starting point for future analysis. The manuscript is written very clearly and the careful selection of figures make it easy to follow the different steps in the analysis.

Therefore, I fully recommend publication of the current manuscript.

I would only ask for minor revision in terms of the description of the methodology in order to ensure the reproducibility of the analysis. First, the description of the pre-processing of the data seems insufficient. For example, when using the NLTK-tokenizer, did the authors filter any words? Second, I couldnt find any information on the repository where the code for the analysis of the data will be published (criteria for publication state that "Datasets, code, and other digital materials should be deposited in an appropriate, recognised, publicly available repository").

Reviewer: 2

Comments to the Author(s)

This manuscript investigates Heaps' law in literary texts. The main contribution of this manuscript is the analysis of the words classified by different parts of speech. As far as I am aware this is the first manuscript that performs this analysis, which adds a meaningful contribution to this area of study. The manuscript is clearly written and the statistical analysis, including the comparison to null models, are correctly performed. I recommend the manuscript for publication after the authors address the points listed below.

1) I found the second sentence of the abstract unclear, and the whole abstract over complicated. I suggest trying to simplify this sentence and to focus on the main results.

2) It would be helpful if the authors would use the distinction between word types and word tokens, which is standard in linguistics. For instance, in the caption of Fig. 2 it is sometimes hard to distinguish which concept is being referred (Also in other parts of the manuscript).

3) The main result of the manuscript in Fig. 2 is very interesting. For large values of N_tag the newly added V_tag must correspond to very rare words, possibly including words that are not in standard dictionaries or list of words. These results rely heavily on the POS tagger and it'd be in general important to add more information about how it is tagging the words. It is remarkable
that the POS tagger seem to consistently tag these words as nouns, verbs, and others. It'd be helpful if the authors would add some information about how the POS tagger works and whether it can be trusted even for extremely rare words? Should the scaling be expected even for N->\infty and is the tagging reliable in this limit?

4) In Sec. 4 and in the third paragraph of the discussion 6 the authors discuss the comparison to randomized texts. This is an important part of the manuscript, which is indeed very relevant and contains original contribution. However, I believe that the observation that randomized texts show larger $V_{\text{tags}}$ is not new. It has been observed and derived mathematically in Ref. 22, but probably it was known even earlier. It is natural to expect taking into account that the usage of words in the text is correlated, with words clustering in regions of the text. This correlation delays the appearance of new words. It is not present in shuffled text, which therefore have a larger number of distinct word tokens for texts of similar length. I believe the authors should revise their claims of novelty in this aspect, specially in the Discussion.

5) All data is available and the Dryad server is working. However, it is not clear how to replicate the results because there is no code available for the data filtering or for the NLP part. The books in the repository still contain metadata, such as translator's preface, and it is not clear how the authors performed the filtering (if any).

Author's Response to Decision Letter for (RSOS-200008.R0)

See Appendix A.

Decision letter (RSOS-200008.R1)

21-Feb-2020

Dear Dr Zanette,

It is a pleasure to accept your manuscript entitled 'Heaps' law and Heaps functions in tagged texts: Evidences of their linguistic relevance" in its current form for publication in Royal Society Open Science. The comments of the reviewer(s) who reviewed your manuscript are included at the foot of this letter.

Please ensure that you send to the editorial office an editable version of your accepted manuscript, and individual files for each figure and table included in your manuscript. You can send these in a zip folder if more convenient. Failure to provide these files may delay the processing of your proof. You may disregard this request if you have already provided these files to the editorial office.

You can expect to receive a proof of your article in the near future. Please contact the editorial office (openscience_proofs@royalsociety.org) and the production office (openscience@royalsociety.org) to let us know if you are likely to be away from e-mail contact -- if you are going to be away, please nominate a co-author (if available) to manage the proofing process, and ensure they are copied into your email to the journal.

Due to rapid publication and an extremely tight schedule, if comments are not received, your paper may experience a delay in publication.
Please see the Royal Society Publishing guidance on how you may share your accepted author manuscript at https://royalsociety.org/journals/ethics-policies/media-embargo/.

Thank you for your fine contribution. On behalf of the Editors of Royal Society Open Science, we look forward to your continued contributions to the Journal.

Kind regards,
Andrew Dunn
Royal Society Open Science Editorial Office
Royal Society Open Science
openscience@royalsociety.org

on behalf of Professor Matjaz Perc (Associate Editor) and Mark Chaplain (Subject Editor)
openscience@royalsociety.org

Follow Royal Society Publishing on Twitter: @RSocPublishing
Follow Royal Society Publishing on Facebook:
https://www.facebook.com/RoyalSocietyPublishing.FanPage/
Read Royal Society Publishing's blog: https://blogs.royalsociety.org/publishing/
Appendix A

RESPONSE TO REVIEWERS - Manuscript ID RSOS-200008

REVIEWER 1

> First, the description of the pre-processing of the data seems insufficient. For example, when using the NLTK-tokenizer, did the authors filter any words?

Response: We are now more specific in section 2 on how the analyzed texts are pre-processed and give some details on the tokenization and tagging algorithms. In particular, tokenization does not use any kind of word filtering. In any case, full documentation of the NLKT library is available in Refs. 25 and 26, and a related reference (now Ref. 27) has been added.

> I couldn't find any information on the repository where the code for the analysis of the data will be published (criteria for publication state that "Datasets, code, and other digital materials should be deposited in an appropriate, recognised, publicly available repository")

Response: Computational codes have now been made available in Dryad.

REVIEWER 2

> 1) I found the second sentence of the abstract unclear, and the whole abstract over complicated. I suggest trying to simplify this sentence and to focus on the main results.

Response: We have rephrased several sentences in the abstract, emphasizing our conclusions. We hope that the text is now clearer.

> 2) It would be helpful if the authors would use the distinction between word types and word tokens, which is standard in linguistics. For instance, in the caption of Fig. 2 it is sometimes hard to distinguish which concept is being referred (Also in other parts of the manuscript).

Response: Following this suggestion, we now use the nomenclature of "word tokens" and "word types" along the text. For the convenience of the reader who may be not familiar with it, we define word tokens and types the first time they appear, in the second paragraph of the introduction.

> 3) It'd be helpful if the authors would add some information about how the POS tagger works and whether it can be trusted even for extremely rare words? Should the scaling be expected even for N->\infty and is the tagging reliable in this limit?

Response: In section 2, we now give more details on the NLTK tagger (see first response to Reviewer 1). Although there are no specifications on how "extremely rare" a word has to be for the method to fail, we have not been able to detect any gross error in the tagging process. Although it might be an interesting question from a formal viewpoint, considering the limit N->\infty makes little sense in our analysis, as unitary English texts spanning more than a few hundred thousand word tokens
(like some of the works considered here, cf. dic09, dic10) are, to our knowledge, inexistent.

> 4) I believe that the observation that randomized texts show larger V_tags is not new. It has been observed and derived mathematically in Ref. 22, but probably it was known even earlier. It is natural to expect taking into account that the usage of words in the text is correlated, with words clustering in regions of the text. This correlation delays the appearance of new words. It is not present in shuffled text, which therefore have a larger number of distinct word tokens for texts of similar length. I believe the authors should revise their claims of novelty in this aspect, specially in the Discussion.

Response: The corpora studied in Ref. 22 consist of collections of untagged texts of different origins, and the analysis of the vocabulary size as a function of text length is not done -as is our case- along each individual text (i.e. as the text progresses) but for the total vocabulary and length of each text (i.e. across texts; see their section 3). However, their discussion on word clustering and topicality is relevant to our comment on the possible causes of retardation in the appearance of new word types in real texts. We have therefore lengthened this comment in section 6, and cited the reference again.

> 5) All data is available and the Dryad server is working. However, it is not clear how to replicate the results because there is no code available for the data filtering or for the NLP part. The books in the repository still contain metadata, such as translator's preface, and it is not clear how the authors performed the filtering (if any).

Response: Computational codes have now been made available in Dryad.