Community proofreading as a tool for community engagement

A quantitative analysis

Sebastian Nordhoff

Enlarging the scope of Open Access to Open Publishing

1 One of the main motivations of Open Access has been to make scientific findings available for everyone to read. This was complemented later by desires to allow everybody to furthermore reuse, redistribute, and adapt the content, in "any digital medium for any responsible purpose, subject to proper attribution of authorship" (Berlin Declaration on Open Access). Open Publishing takes this further as it pushes for open availability of workflows, business models, business figures and analytical insights (Caux 2017; Nordhoff 2018a). This paper shares the insights of organising proofreading/copy-editing in a community-based fashion. It will specify the organisational setup, the software implementation, and provide quantitative data for evaluation.

Community

2 Bibliodiversity is defined as a complex system of various actors which assume various roles. They might be readers on Monday, authors on Tuesday, and reviewers on Wednesday. More junior researchers might more often assume the reader's role, taking up the other roles more frequently as their career progresses. A community-based publisher relies on this ecosystem and tries to integrate researchers at different stages of their career, finding tasks for everybody at their respective level. The inclusion of junior researchers empowers them on the one hand, and leads to a higher identification with the community-based publisher on the other.

3 For new publishers, community engagement is an important task. This paper will show how crowdsourcing the proofreading process can be used to engage the community and make them adapt the new platform as « their » platform.

Language Science Press

4 Language Science Press is a community-based publisher, which has published more than 100 open access linguistics books since 2014. Language Science Press is committed to openness. Next to the pdfs, Language Science Press shares the LaTeX source code of all books, the original graphics in high resolution, all software they produce, their workflows, their business model, and most of their business figures (Nordhoff 2018ab).

LangSci workflow

5 Language Science Press has a collaborative workflow, with a records-of-versions approach. The following versions are distinguished:

1. Submission version
2. (Optional Open Review version)
3. Community proofreading version
4. First edition
5. Subsequent editions

6 All versions after the Open Review version are publicly available for inspection and commenting on PaperHive. PaperHive is an online annotation platform where registered users can select passages and comment on them, similar to hypothes.is. This paper will focus on the way from the community proofreading version to the first edition (steps 3 and 4).

Community proofreading

7 The traditional setup in linguistics publishing is that the publisher hires a copy-editor to go over the book. The copy-editor will send an annotated document to the author. Annotations focus on language and the application of the publisher's style sheets. Copy-editors may or may not be specialists in the particular subfield the book covers. Most copy-editors probably have special training for their task, but this is normally not disclosed by the publishers. Copy-editing is work-for-hire.

8 Community proofreading takes a different approach: crowd-sourcing. Instead of recruiting a specialist in grammar and style, the manuscript to be published is offered to a pool of interested linguists in its final draft form for "sneak preview." Everyone who is interested can comment. Proofreaders get early access to new research in return for commenting on typos, errors or possible misunderstandings. Since the population of proofreaders is very similar to the population of actual readers of the final book, the issues they mention are the ones which are indeed the ones which might confuse the final readership. They might not spot all Oxford commas, but they will spot sentences or passages which are hard to understand, even if, from a purely technical point of view, they respect all guidelines and manuals.

9 Community proofreading is voluntary work. As such, only small portions of a book are assigned to every proofreader (typically 1 chapter). For a book with 8 chapters, the aim would be to have 16 proofreaders (2 per chapter). The redundancy is needed for drop-outs or to compensate for low quality.

10 Language Science Press has built up a pool of 350 proofreaders. Every other week, a new title is announced to the proofreaders list. Interested community members then have 2 days to volunteer and claim a chapter. The coordinator assigns the chapter and takes care that chapters are covered equally, and that the thematic wishes of the volunteers are respected.

11 All proofreaders work on PaperHive, where they annotate the same document. The proofreading phase lasts 4 weeks. After 4 weeks, proofreading is closed and the comments are made available to the author.

Study

12 After the description of the setup, let's turn to evaluation. We can distinguish two types of areas to evaluate: the quantity and quality of annotations, and the question whether community proofreading leads to a higher engagement with the press. A first analysis of community proofreading has been provided by Westedt (2018). She looked at a sample of book chapters and categorised all comments she found into grammar, style, references, content and so on. One basic finding of hers is that community proofreading actually goes beyond traditional copy-editing in that many comments pertain to content issues on a very specialised level (Table 1).

Table 1

Category	Percentage
Spelling	7.3
Syntax	7.8
Lexical choice	20.73
Grammar	11.55
Punctuation	11.81
Style	21
Content	6.56
Miscellanea	3.41
References	9.71

13 This paper expands Westedt's findings with a study based on a larger corpus of 52 books. While Westedt did a qualitative assessment of each comment and categorised it, in this

paper, I take a purely computational approach, which is mandated by the size of the corpus.

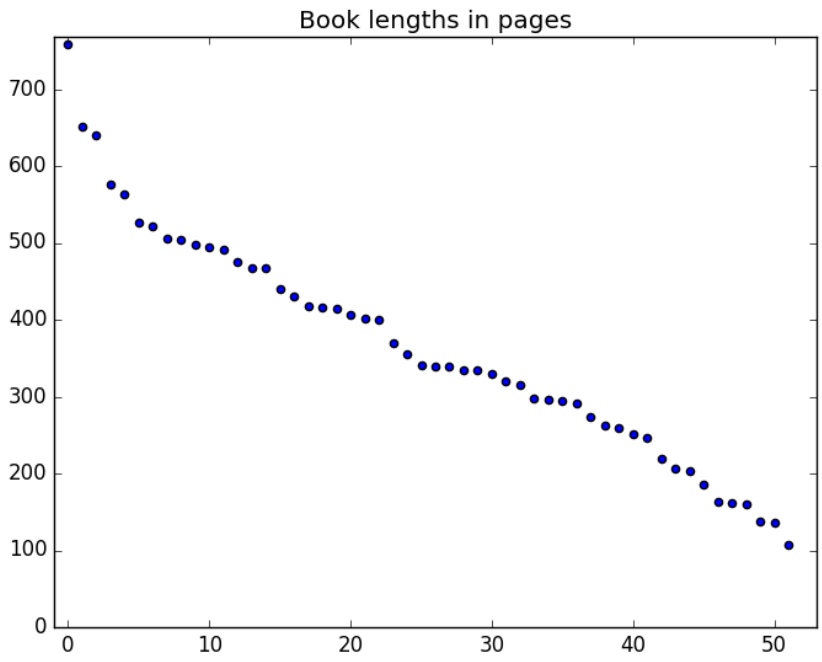
The corpus

14 The corpus covers 52 books which entered the proofreading phase after November 2016 with a total of 19,004 pages. Of those, 10,388 pages have at least one comment, for a total of 43,370 comments. Only works in English are considered for this study. The corpus including scripts and graphics is available from http://doi.org/10.5281/zenodo.3063004.

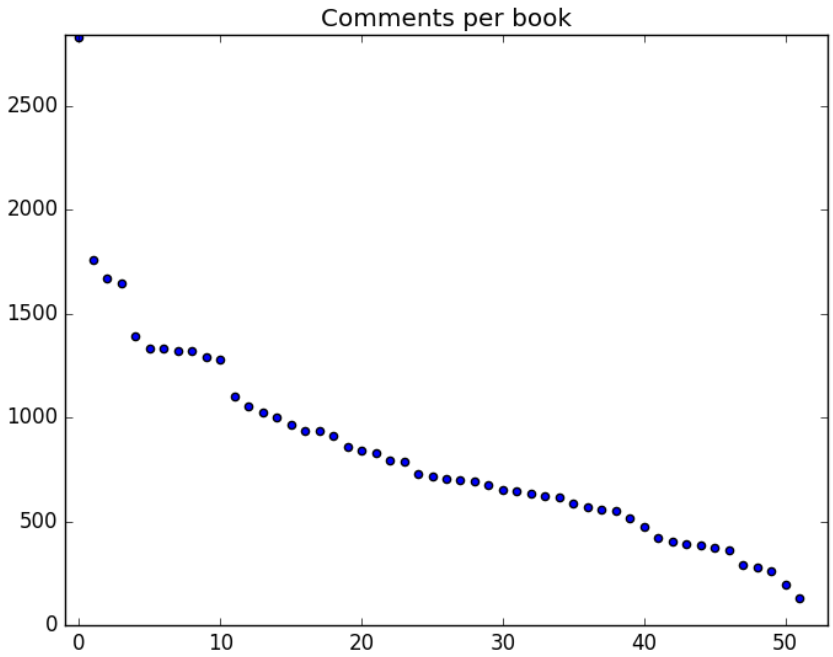
Descriptive statistics

15 Books vary in length between 100 and over 700 pages (Plot 1). The number of comments per book goes from 100 to over 2700 (Plot 2). The average number of comments per page per book is given in Plot 3. The highest number of comments on one page is found in *Theory and description in African Linguistics* on page 122[1] (48 comments).

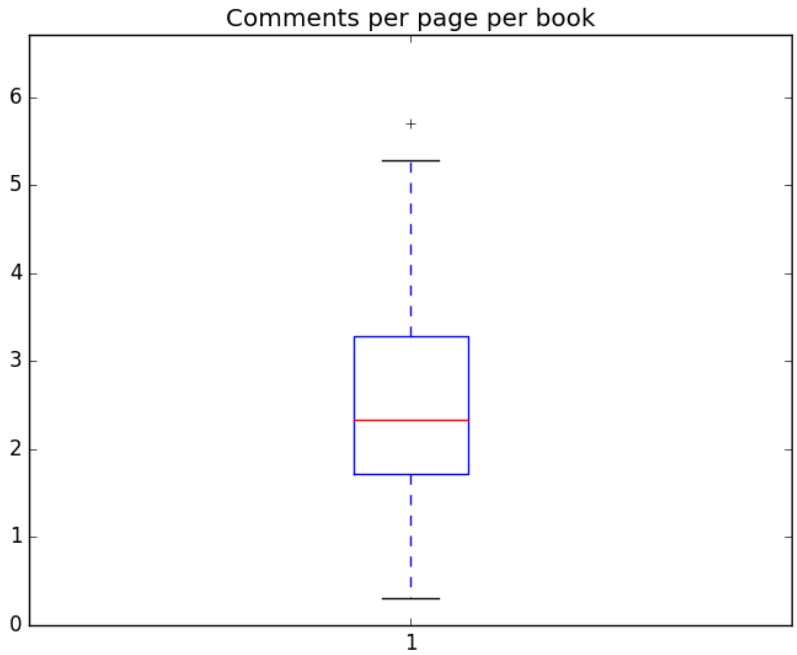
Plot 1



Plot 2



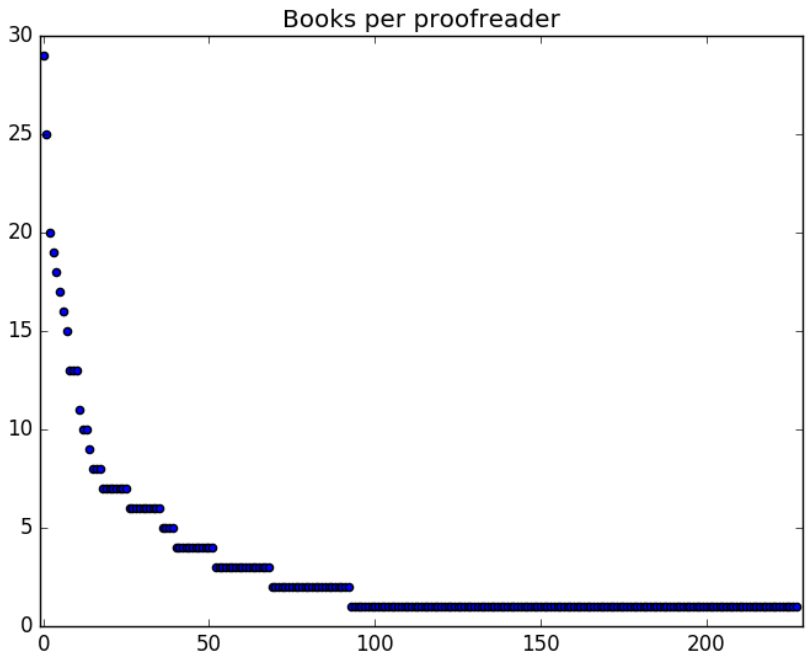
Plot 3



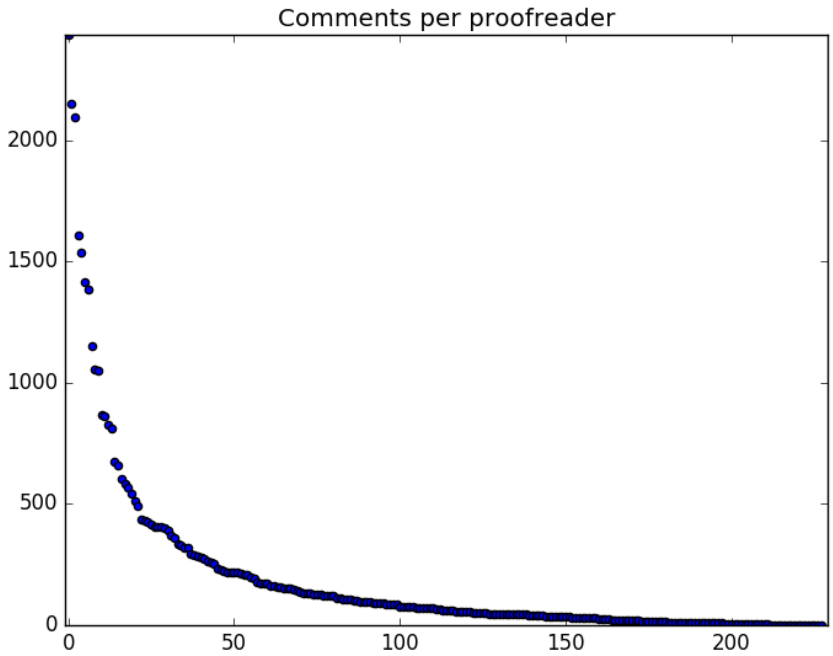
16 228 different accounts have participated in commenting. Some proofreaders have been more active than others: the top proofreaders have participated in more than 25 books,

and have left more than 2000 comments (Plots 4 and 5). Conversely, the number of proofreaders per book is typically around ten, but can reach over 40 on occasion (Plot 6).

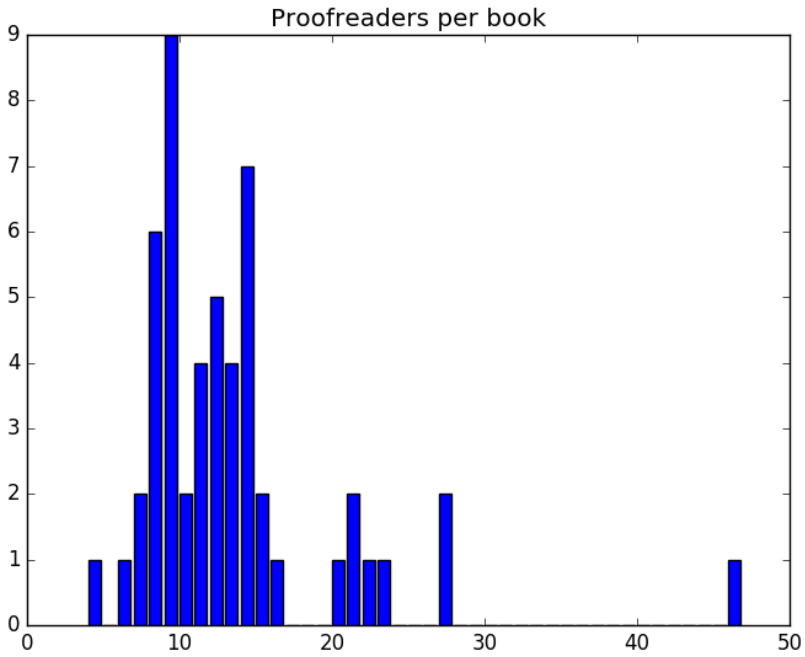
Plot 4



Plot 5

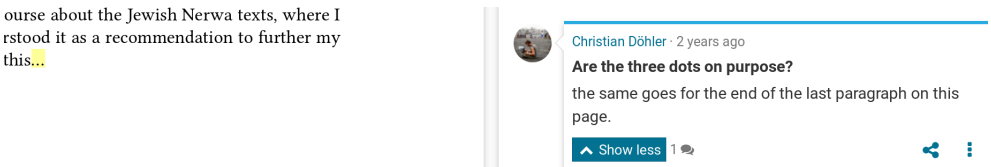


Plot 6



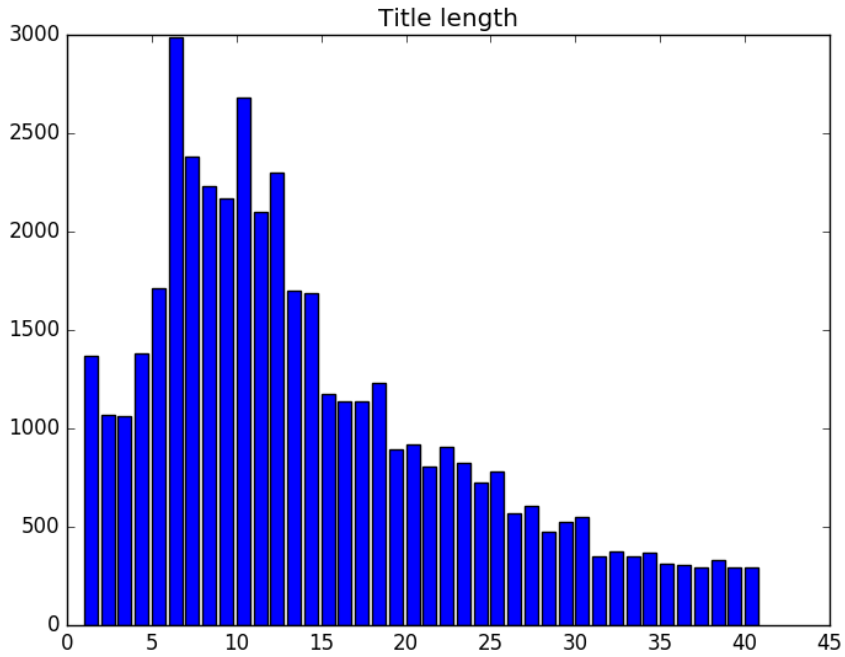
17 On PaperHive, a comment has a succinct title (such as “Are the three dots on purpose?” Figure 1) and can optionally have a body, where more elaborate information as to the issue at hand is given (such as “the same goes for the end of the last paragraph on this page,” Figure 1).

Figure 1

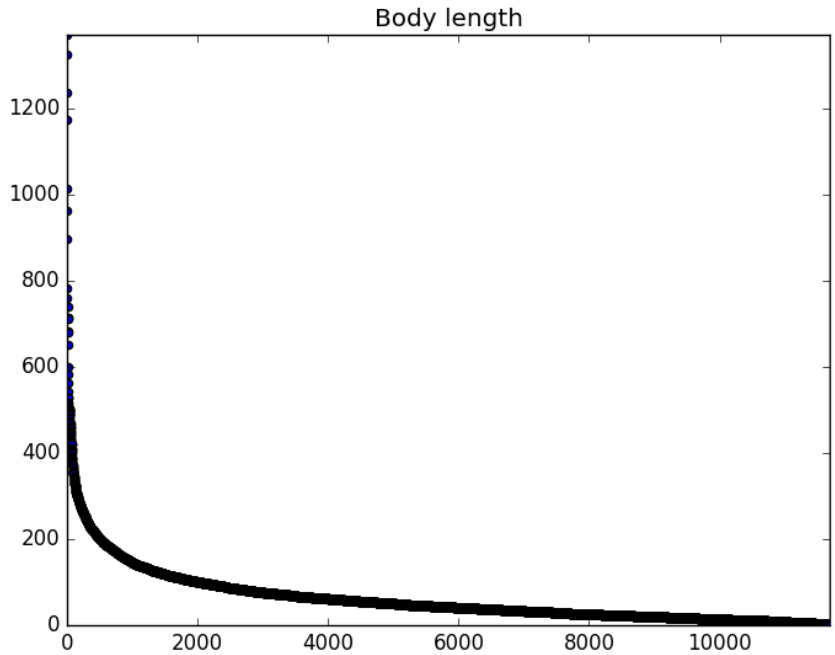


18 Titles cannot have more than 40 characters. It turns out that most titles only need about 15 characters, so that the 40 characters limit does not present a problem for standard cases (Plot 7). The body of the comment is typically empty, but if more explanation is given, it can become very large (over 1200 characters; Plots 8 and 9).

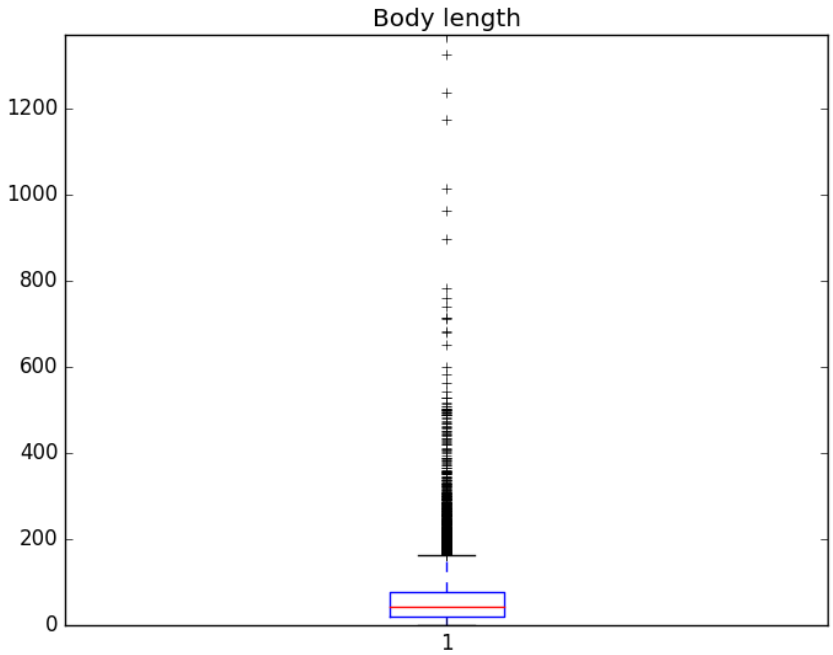
Plot 7


Title length
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
35
40
45

Plot 8


Body length
1200
1000
800
600
400
200
0
0
2000
4000
6000
8000
10000

Plot 9



Creation of the corpus

19 The script paperhive2tsv.py distributed together with this document allows to retrieve all comments for a LangSci book if the ID (e.g. 144) is known. The comments are downloaded and stored as tab-separated values in one file per book.[2] The individual tsv-files can be concatenated to form a global file. The script tsv2sqlite.sh loads this file into an SQLite database.[3] The script analyzeCPR.py then uses this database for its queries and further processing, as shown in the following section.

Quantitative hypotheses

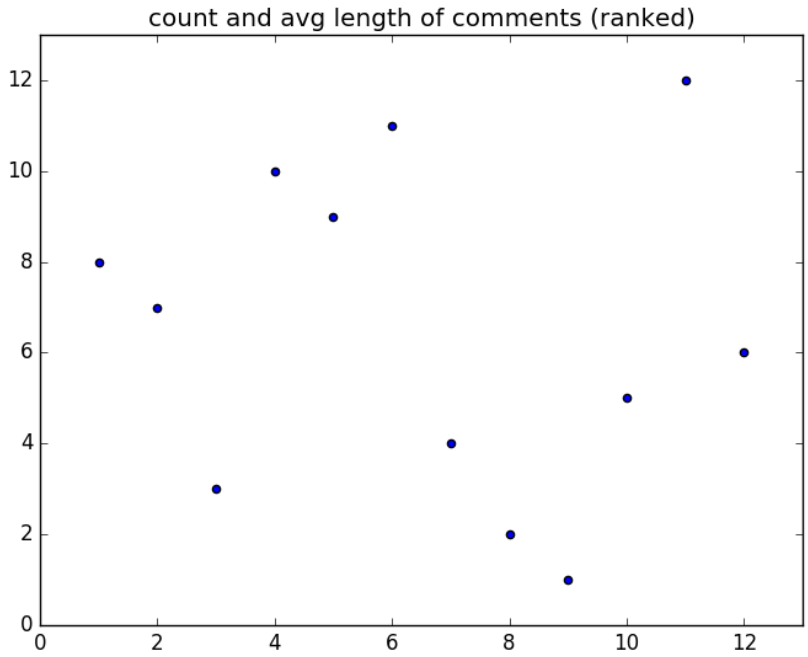
20 Next to purely descriptive statistics, the proofreading corpus can also be used to test hypotheses about how proofreaders interact with a text. I have tested two such hypotheses:

Hypothesis 1: Proofreaders fall into two types. Type 1 will focus on small details; type 2 will focus on the big picture

21 Type 1 should have lots of short comments ("comma missing"), while Type 2 should have fewer comments, which would however be more elaborate. In order to test this hypothesis, we can, for every book, establish the number of comments per proofreader and the average comment length per proofreader. We then establish two rankings: R1 will give the order of proofreaders according to number of comments while R2 will give the order of proofreaders according to comment length. For every book, we can plot the two rankings against each other. An example from *Empirical modelling of translation and*

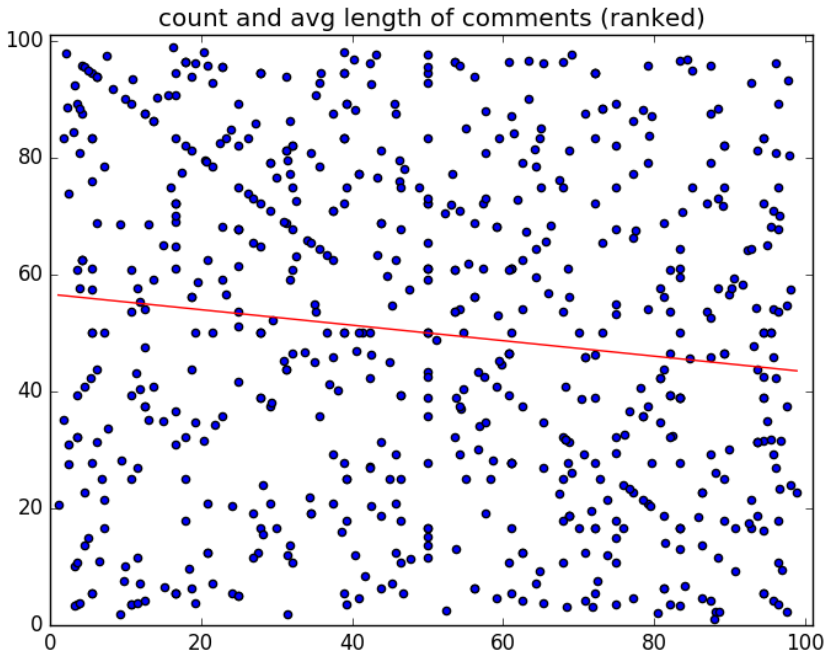
interpreting is given in Plot 10. 12 proofreaders have participated in this book and their respective ranks are given by the dots.

Plot 10



22 We see that the proofreader ranked #3 in one domain is also #3 in the other, but the proofreader ranked #1 in one domain is only #8 in the other. This book in itself does not give us sufficient evidence to confirm our hypothesis. The hypothesis would have predicted that the dots are distributed along a straight line. In order to broaden the empirical base, we can have a look at all books. To make the ranks comparable across books with different amounts of proofreaders, the ranks are normalized to 1..100. We can then superpose all plots. The plot including all books is given in Plot 11.

Plot 11



23 The result is still messy, but we can compute a best fit, which is given by the red line. This line shows that there is a slightly negative correlation between the two rankings: the higher you are in one rank, the lower will you be in the other. The hypothesis is therefore confirmed.

Hypothesis 2: Proofreading comments will diminish as the proofreader moves along. Comments will become fewer due to fatigue, and average comment length will go down due to repetition of previous remarks as "see above"

24 In order to test this hypothesis, we have to establish the relative length and the relative position of comments for each proofreader and each book. Are they more towards the start, the middle or the end of a chapter?

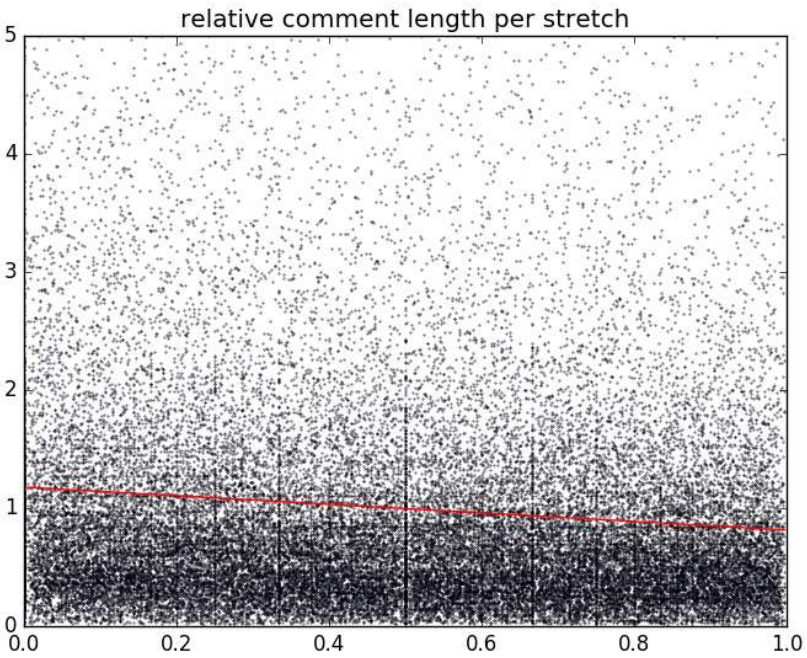
25 The relative length can easily be computed as the relation of the length of a given comment to the average length of all comments considered (e.g. 137%).

26 The relative position of a comment can be established in two fashions: either take the sequence of comments as the base, or the page number. In the first approach, we enumerate all comments. Say there are 11 comments. Comment #6 will then be exactly in the middle, since there are 5 preceding comments and 5 following comments. In the second approach, we look at the page numbers of the relevant stretch.[4] Suppose the eleven comments are on pages [101, 103, 104, 105, 107, 110, 111, 120, 121, 130, 140].[5] Comment #6 will be on page 110, which happens to be the tenth page out of 40 pages under consideration, or at 25% down the chapter (compare this to the first approach, where comment #6 was at 50%).

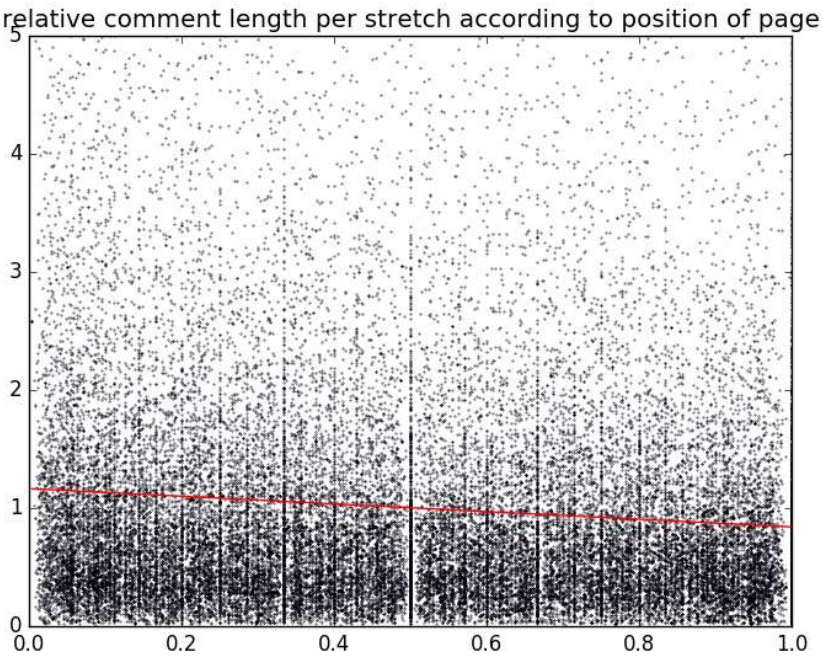
27 We can now plot the relative length of a comment against its relative position. The prediction is that a later position in a stretch will lead to a shorter relative length. The

plots are given in Plot 12 for the comment-sequence-based approach and in Plot 13 for the page number based approach. A dot at (0.5, 5) means that there was a comment in the middle of the relevant stretch whose length was 5 times the average comment length.

Plot 12



Plot 13



28 Both approaches lead to comparable results: The first comment in a chapter is likely to have a length of about 110% of the average, while the last comment is likely to be about 90% of the average. The hypothesis of "proofreader fatigue" is thereby confirmed. The effect is not strong, but discernible.

29 This paper is a methodological paper, which outlines the setup for community proofreading and shows how it interacts and integrates with Open Publishing. The two hypotheses explored here do not strictly pertain to the research questions which are commonly asked in the context of Open Access, but they show that once the documents, processes, and formats are opened up, novel research questions can emerge which would not have been possible under a closed setup.

The ecosystem

30 The study of comments is complemented by a study of shifting roles within the LangSci ecosystem. Are the sets of proofreaders and authors disjunct, or are there researchers which form part of both groups? If so, is it the case that proofreaders become authors, or is it the case that authors want to return the proofreading services they received and become proofreaders?

31 I compiled a list of all "producers" (volume editors chapter authors, monograph authors). This list features 908 different names. I matched this against the list of 188 proofreaders in the LangSci Hall of Fame (http://langsci-press.org/hallOfFame).[6] The intersection of the two lists has 27 names. Of those, 11 started as authors and took up proofreading later, whereas 16 started as authors and later took on some proofreading tasks as well. This means that there is movement between the author pool and the proofreader pool in both directions.

Conclusions

32 Community proofreading is a novel way of engaging the community, inscribed in the context of Open Publishing. It is only possible for Open Access publications since there is no need to artificially restrict access to content in order to monetise it. The implementation has proved to be workable and has been used by more than 200 researchers. Qualitative findings suggest that community proofreading can compare to traditional forms of proofreading, and covers similar areas. As a by-product, data about how proofreaders interact with a text are generated, which can be useful for studying the process of reading and text comprehension, e.g. in the fields of psychology, pedagogy, or library science. This has been shown above for the idea of "proofreader fatigue."

33 While community proofreading can compare to traditional copy-editing, it is not simply a cheapie substitute. It has its own strengths (content) and weaknesses (consistency), and, depending on a particular use case, one or the other approach might be more appropriate. Given the open nature of the process, the data is available for quantitative analysis, allowing other publishing projects to profit from the LangSci experiences. Quantitative data suggests that good community proofreading can achieve good coverage, and that different types of proofreaders can be distinguished. The data furthermore show that there is a flow back and forth between the group of authors and the group of proofreaders, indicating a healthy ecosystem where researchers from

different backgrounds at different stages of their career contribute their respective expertises to creating and improving manuscripts.

BIBLIOGRAPHY

References

Nordhoff, Sebastian. 2018. *Cookbook for Open Access books*. Berlin: Language Science Press. http://doi.org/10.5281/zenodo.1286925

Nordhoff, Sebastian. 2018. *Language Science Press business model*. Berlin: Language Science Press. http://doi.org/10.5281/zenodo.1286972

Westedt, Lole. 2018. "Community Proofreading am Beispiel Language Science Press: „Gratis-Korrekturlesen" oder auch inhaltlich anreichernd?" Unpublished BA thesis, Humboldt Universität Berlin. https://paperhive.org/documents/items/Tjjj5pQKj2ci

Caux, Jean-Sébastien. 2017. "Noble metals for a noble cause." September 20. https://jscaux.org/blog/post/2017/09/20/noble-metals-noble-cause/

Ross-Hellauer Tony. 2017. "What is open peer review? A systematic review. [version 2; peer review: 4 approved]." *F1000Research* **6**:588. https://doi.org/10.12688/f1000research.11369.2

NOTES

1. https://paperhive.org/documents/items/tIIak2Ks9tEy?a=p:122
2. For books with an Open Review version, the PaperHive ID cannot be retrieved automatically, but must be supplied manually.
3. The extra step via *tsv (rather than direct import into the database) was taken to facilitate reuse by other projects.
4. There is a complication in that proofreaders are sometimes assigned non-adjacent chapters. To take care of this, a passage of more than 20 intervening pages without comments between two comments by the same proofreader is taken to establish that the two stretches form part of different chapters.
5. There can be more than one comment on one page, but for the sake of exposition, this is ignored here as it does not affect the results.
6. The difference to the number of proofreaders on PaperHive is due to a) people having opted out of the Hall of Fame and b) the Hall of Fame only considering books which are already published whereas for the PaperHive study above, all books where proofreading was finished were considered, even if the book is not published yet.

ABSTRACT

This paper describes Community Proofreading as implemented by Language Science Press via PaperHive. Community members comment on a final draft version of a book and highlight possible improvements. A database of over 43.000 comments was compiled, which allows for the formulation of novel research questions. Two of those ("small details vs. big picture" and "reviewer fatigue") are tested in this paper. Furthermore, the paper shows that Community Proofreading can serve as a tool to attract new authors.

INDEX

Keywords: Community Proofreading, Community Engagement, PaperHive, Evaluation

AUTHOR

SEBASTIAN NORDHOFF
Language Science Press
sebastian.nordhoff@langsci-press.org