



Community proofreading as a tool for community engagement

A quantitative analysis

Sebastian Nordhoff

June 4, 2019

FU Berlin

- › Open Access is mainly concerned with reading
- › Open Publishing is concerned with making all aspects of publishing open (Rob Cartolano)
 - › Open source platforms
 - › Open formats
 - › Open protocols
 - › Open bookkeeping
 - › Open peer review
 - › Community proofreading

- › one research can adopt different roles
 - › author, reviewer, reader, ...
- › junior researchers are more often readers
- › senior researchers take on the other roles as well
- › complex ecosystem
- › community-based publishing tries to integrate researchers at all levels

- › outsourced work-for-hire
- › for a fee
- › one proofreader
- › specialist in style and guidelines
- › might have some training in linguistics
- › normally no specialist knowledge of the particular subfield

- › crowdsourced to the community
- › voluntary work
- › many proofreaders, often junior
- › very often specialists in the particular subfield
- › intrinsic interest
- › less acquaintance with style and guidelines

- › Open Access publisher in linguistics
- › 100+ books since 2014
- › 350 community proofreaders

<p>Natural causes of language: Frames, biases, and cultural transmission N. Enfield (Author)</p>	<p>A topology of marked languages Corina Fiebert (Author)</p>	<p>The Talking Heads experiment: Origins of words and meanings L. L. Lewis (Author)</p>	<p>Grammaticalization in the North: Noun phrase morphemes in Scandinavian vernaculars Ole Dahl (Author)</p>	<p>A grammar of Fula (Sami) Jonas Wihor (Author)</p>	<p>Phonetic detail in Neogutierrezian Francesco Cengini (Author)</p>
<p>Linguistic variation, identity construction and cognition Karin K. Dräger (Author)</p>	<p>Language strategies for the domain of labor Jonas Wihor (Author)</p>	<p>A grammar of Fula (Sami) Ole Dahl (Author)</p>	<p>Thoughts on grammaticalization Ole Dahl (Author)</p>	<p>How mobile robots can self-organize a vocabulary Paul Hager (Author)</p>	<p>A grammar of Mawé Luis Berghall (Author)</p>
<p>Roots of language David Beckson (Author)</p>	<p>The empirical base of linguistics: Grammatical judgments and linguistic methodology Carsten T. Schuler (Author)</p>	<p>A grammar of Fula (Sami) Ole Dahl (Author)</p>	<p>New directions in corpus-based translation studies Christy Farnsworth, Frederic Zeman (Volume Editor)</p>	<p>Grammatical theory: From transformational grammar to construction-based approaches Stefan Müller (Author)</p>	<p>Advances in the study of Slavic languages and linguistics Catherine Ruy, Björn Petersen (Volume Editor)</p>
<p>The evolution of grounded spatial language Michael Spranger (Author)</p>	<p>The evolution of case grammar Bernard Trapp (Author)</p>	<p>The future of dialects: Selected papers from Methods in Dialectology IV Marie-Hélène Côté, Bernice Knoch, John Nerbonne (Volume Editor)</p>	<p>Syntax und Valenz: Der Modellierung von Verbstrukturalen und affektiven Strukturen im Baumadjunktogramm Thomas Lohr (Author)</p>	<p>Adjective attribution Michael Heller (Author)</p>	<p>A grammar of Kanyemba Anna Steinhilber (Author)</p>
<p>Sprachliche Sozialisation: Jüdischkeit in der deutschsprachigen Literatur (18.-20. Jahrhundert) Luis Scheller (Author)</p>	<p>Forthcoming: Language technologies for environmental change Gregory Bateson, David G. Bateson, John S. Bateson (Volume Editor)</p>	<p>Eyetracking and Applied Linguistics Silvia Hansen-Schmidt, Sandra Grunz (Volume Editor)</p>	<p>Forthcoming: Grammatical structures in comparative phrase structure analysis Gabriela Wille (Author)</p>	<p>Einführung in die grammatische Beschreibung der Deutschen: Zweite, überarbeitete Auflage Roland Schüller (Author)</p>	<p>New perspectives on cultural coherence: Implications for translation Klaus-Michael Quasthoff, Barbara Krenn (Volume Editor)</p>
<p>Time in Yenching: A lexical study and morphological analysis Alexa Michael (Author)</p>	<p>A grammar of Kanyemba Anna Steinhilber (Author)</p>	<p>Forthcoming: Attribution construction in North-Eastern Neo-Aramaic Joni Gutman (Author)</p>	<p>Forthcoming: Empirical modelling of translation and interpreting Silvia Hansen-Schmidt, Oliver Gals, Sascha Hoffmann, Bernd Meyer (Volume Editor)</p>	<p>Forthcoming: Reflections on Bilingualism Gerd Bacher (Author)</p>	<p>Diversity in African languages: Selected papers from the 46th Conference on African Linguistics Doris L. Payne, Sara P. Harnett, Bruce Vandevoort (Volume Editor)</p>
<p>Annotation, exploitation and evaluation of parallel corpora: TC3 Silvia Hansen-Schmidt, Sandra Grunz, Oliver Gals (Volume Editor)</p>	<p>Forthcoming: Beiträge zur deutschen Grammatik Tobias H. Hilde (Author), Stefan Müller, Frank Richter, Margarete (Volume Editor)</p>	<p>The Bilingual Dictionary and grammar sketch Tobias H. Hilde (Author)</p>	<p>A grammar of Pagan Malay Angela Hager (Author)</p>	<p>Forthcoming: Unity and diversity in grammaticalization scenarios Walter Stroh, Achim Machobane (Volume Editor)</p>	<p>Tonal placement in Tachikawa Tachikawa (Volume Editor)</p>
<p>Forthcoming: Order and structure in syntax II: Subordinate and argument structure Laura A. Bailey, Michelle Sheehan (Volume Editor)</p>	<p>Forthcoming: Order and structure in syntax I: Word order and syntactic structure Laura A. Bailey, Michelle Sheehan (Volume Editor)</p>	<p>Forthcoming: A equívoco da língua materna em português: Questões gerais e específicas da Português Marta João Freitas, Ana Lúcia de Santos (Volume Editor)</p>	<p>Forthcoming: Further investigations in the syntax of infinitives Cristina Tiberius (Volume Editor)</p>	<p>Dependencies in language: On the central argument of linguistic systems N. J. Enfield (Volume Editor)</p>	<p>On looking towards (and beyond) the future Doris L. Payne, Sara P. Harnett, Bruce Vandevoort (Volume Editor)</p>

- › proofreading queue with a new title every 2 weeks
- › title is announced on Monday
- › community members can volunteer and claim a chapter
- › chapters are assigned on Wednesday
- › 4 weeks time for proofreading
- › proofreading is done on Paperhive

Document

Discussions 317

Activity

which would yield only default agreement on the RC predicate, contrary to fact.

2.2 The Genitive of Quantification phenomenon

The Genitive of Quantification phenomenon has been described to a large extent for Slavic languages in Bošković (2006); Franks (1994; 2002); Przepiórkowski (2004); Rutkowski (2002); and Willim (2003), to name but a few. In Polish, genitive case marking is forced on a noun which is modified by a higher numeral or a lower virile numeral, as well as by certain quantifiers such as *wiele* 'many', *kilka* 'a few', *para* 'a couple of', etc. Such numeral phrases do not induce subject-verb agreement in main clauses, as can be seen in (17), in which the verb obligatorily appears in the 3SG neuter form, regardless of the grammatical gender of the noun.

- (17) a. Siedmiu mężczyzn weszło/*weszli do domu.
seven.ACC men.GEN,VIR entered.3SG,NEUT/3PL,VIR into house
'Seven men entered the house.'
- b. Siedem kobiet weszło/*weszły do domu.
seven.ACC women.GEN, NON-VIR entered.3SG,NEUT/*3PL,NON-VIR into house
'Seven women entered the house.'

The analysis of Polish GoQ structures proposed in Witkoś & Dziubała-Szrejmbrowska (2016) follows the idea that probing for phi-features is possible for T

x



Alexandr Rosen · 18 days ago

N instead of NEUT

according to

<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>

Show more



Alexandr Rosen · 18 days ago

GEN.VIR

period as the separator
according to

<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>

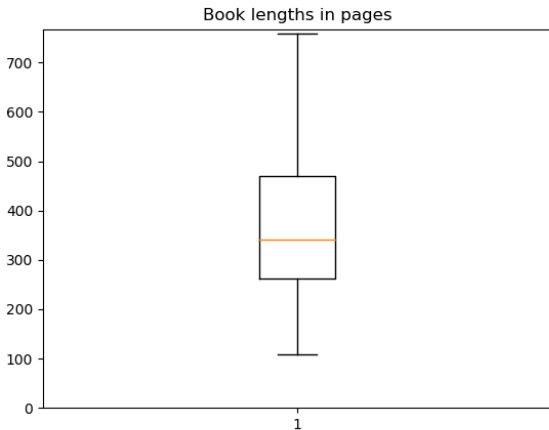
Show more

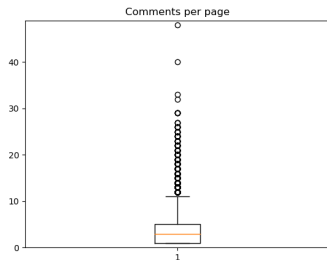
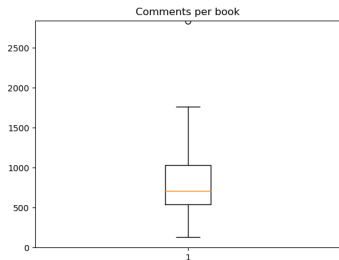


} Westedt analysed a sample of comments on Paperhive for her BA thesis.

Category	Percentage
Style	21.00
Lexical choice	20.73
Punctuation	11.81
Grammar	11.55
References	9.71
Syntax	7.80
Spelling	7.30
Content	6.56
Miscellanea	3.41

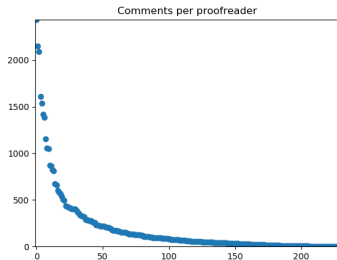
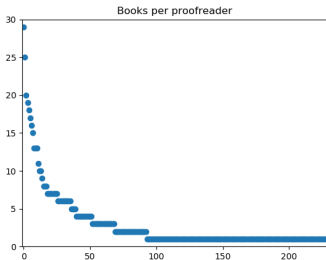
- › 52 books from late 2016 to late 2018
- › comments were harvested from Paperhive and put into a database
- › 19 004 pages
- › 43 370 comments
- › data on <https://doi.org/10.5281/zenodo.3063004>



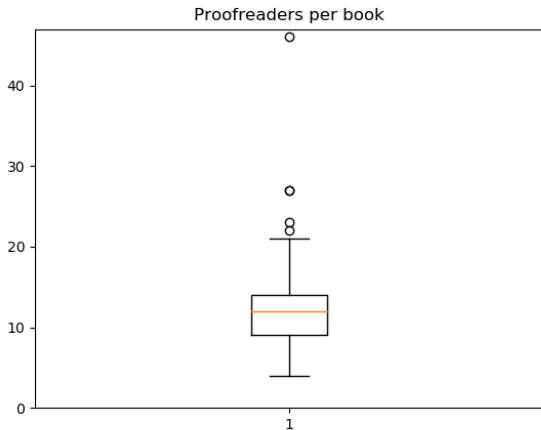


The highest number of comments on one page is found in Theory and description in African Linguistics on page 122 (48 comments).

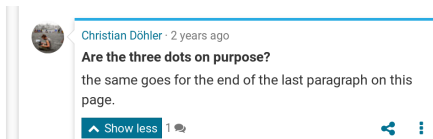
Productivity of proofreaders



228 different accounts have participated in commenting.

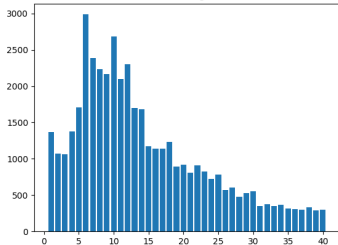


course about the Jewish Nerwa texts, where I
rstood it as a recommendation to further my
this...

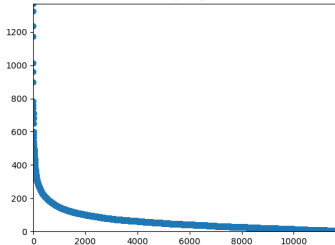


- › A PaperHive comment has a succinct title (<40 characters)
- › optional body, with more elaborate information

Title length



Body length

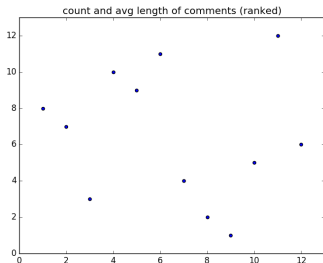


1. **Proofreaders fall into two types. Type 1 will focus on small details; type 2 will focus on the big picture.**
2. **Proofreading will diminish as the proofreader moves along. Comments will become shorter due to fatigue, i.e. average comment length will go down due to repetition of previous remarks as “see above”.**

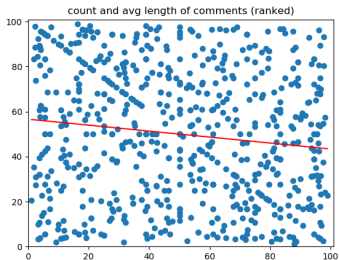
- › Type 1: many comments but short (“comma missing”)
- › Type 2: few comments, but longer, in-depth

› For every book

- › rank all participating proofreaders by amount of comments
- › rank all participating proofreaders by average length of comments
- › plot the two against each other



- > 12 proofreaders participated
- > their respective ranks are given by the dots.
 - > e.g. #3 in one rank is also #3 in the other, but #1 on one is #8 in the other
- > data from one book insufficient



- > Ranks are normalized to centiles
- > best fit given by red line
- > indeed a weak negative correlation

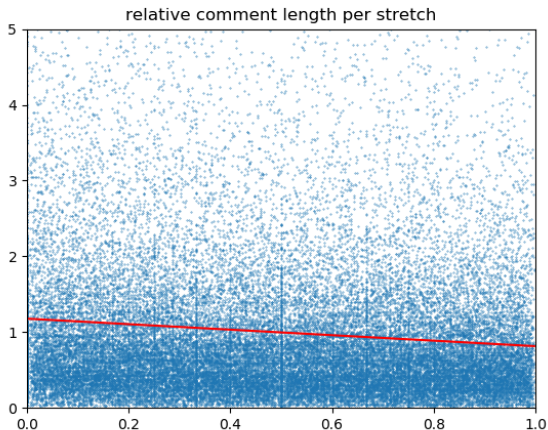
} Hypothesis #1 is confirmed

- } proofreaders with more comments have shorter comments
- } proofreaders with longer comments comment less

Hypothesis 2 : Proofreading will diminish as the proofreader moves along. Comments will become shorter due to fatigue, i.e. average comment length will go down due to repetition of previous remarks as “see above”.

- › for every book
 - for every proofreader
 - for every comment
 - › compute relative length (e.g. 0.67 of the average)
 - › compute relative position (front, middle, back)
 - › store the tuple (relative position, relative length)
 - › 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.
- › the relative position can be pegged to the linear order of comments, or to the pages

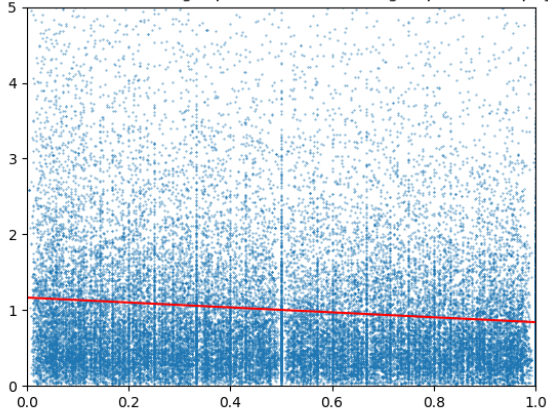
Plot for Hypothesis #2 based on linear order



Plot for Hypothesis #2 based on page position

| Hypothesis evaluation
| Proofreader fatigue

relative comment length per stretch according to position of page



› Hypothesis is confirmed

- › the later in the document a comment is, the shorter it will be
- › the first comment will be about 110% of the average, while the last one will be 90% of the average.
- › effect not very strong, but discernible

- › Main aim: methodological
- › Proofreading comments are a by-product of open publishing
 - › In traditional publishing models, these data would not be available
- › Once the documents, processes, and formats are opened up, novel research questions can emerge which would not have been possible under a closed setup.
- › Implications for psychology of reading for instance.

- › There are 908 people with the role “author” at LangSci Press
- › There are 228 proofreaders
- › 27 researchers have taken up both roles
 - › 16 started as authors, and became proofreaders later
 - › 11 started as proofreaders, and became authors later
 - › Movement between the author pool and the proofreader pool in both directions.

- › Community proofreading is a novel way of engaging the community
- › only possible for Open Access publications
- › workable implementation with 50+ books and 200+ researchers
- › can compare to traditional proofreading
- › by-product data can be used for novel research questions
 - › proofreader typology
 - › proofreader fatigue
- › flow back and forth between the group of authors and the group of proofreaders
- › healthy ecosystem
- › researchers from different backgrounds at different stages of their career contribute their respective expertises to creating and improving manuscripts.

- › What other questions could be addressed with that data?
- › Which other disciplines might be interested?



Gold proofreaders

› **Andreas Hölzl** ([view profile](#))



34/99

› **Jeroen van de Weijer** ([view profile](#))



31/99

› **Eitan Grossman** ([view profile](#))

29/99

› **Jean Nitzke**

25/98

› **Christian Döhler** ([view profile](#))

19/97

› **Martin Haspelmath** ([view profile](#))

19/97

› **Ahmet Bilal Özdemir** ([view profile](#))

19/97

› **Ikmi Nur Oktavianti** ([view profile](#))

15/96

› **Brett Reynolds** ([view profile](#))

14/96