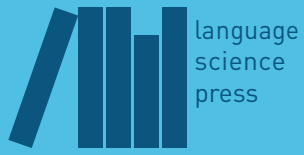
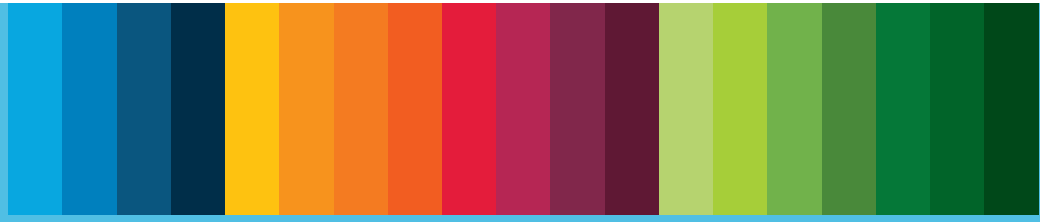




Community proofreading as a tool for community engagement

A quantitative analysis

Sebastian Nordhoff

June 3, 2019

FU Berlin

Open Publishing

- 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

Bibliodiversity

- 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

Traditional proofreading

- 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

Community proofreading

- 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

Language Science Press

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

Books


Natural causes of language: Frames, biases, and cultural transmission
N.J. Enfield (Author)
A typology of marked-S languages
Corinna Handschuh (Author)
The Talking Heads experiment: Origins of words and meanings
Luc L. Steels (Author)
Grammaticalization in the North: Noun phrase morphosyntax in Scandinavian vernaculars
Östen Dahl (Author)
A grammar of Pite Saami
Joshua Wilbur (Author)
Prosodic detail in Neapolit
Francesco Cangemi (Author)
Linguistic variation, identity construction and cognition
Katie K. Drager (Author)
Language strategies for the domain of colour
Joris Bleys (Author)
A grammar of Yakkha
Diana Schackow (Author)
Thoughts on grammaticalization
Christian Lehmann (Author)
How mobile robots can self-organise a vocabulary
Paul Vogt (Author)
A grammar of Mauwake
Lisa Berghäll (Author)
Roots of language
Derek Bickerton (Author)
The empirical base of linguistics: Grammaticality judgments and linguistic methodology
Carson T. Schütze (Author)
A grammar of Palula
Henrik Liljegren (Author)
New directions in corpus-based translation studies
Claudio Fantinuoli, Federico Zanettin (Volume Editor)
Grammatical theory: From transformational grammar to constraint-based approaches
Stefan Müller (Author)
Advances in the study of Si
languages and linguistics
Catherine Rudin, Bryan James G
(Volume Editor)
The evolution of grounded spatial language
Michael Spranger (Author)
The evolution of case grammar
Remi van Trijp (Author)
The future of dialects: Selected papers from Methods in Dialectology XV
Marie-Hélène Côté, Remco Knooihuizen, John Nerbonne (Volume Editor)
Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken
Timm Lichte (Author)
Adjective attribution
Michael Rießler (Author)
A grammar of Yauyos Quec
Aviva Shimelman (Author)
Sprachliche Imitation: Jiddisch in der deutschsprachigen Literatur (18.–20. Jahrhundert)
Lea Schäfer (Author)
Forthcoming: Language technologies for a multilingual Europe
Georg Rehm, Daniel Stein, Felix Sasaki, Andreas Witt (Volume Editor)
Eyetracking and Applied Linguistics
Silvia Hansen-Schirra, Sambor Grucza (Volume Editor)
Forthcoming: Grammaire des constructions elliptiques: Une étude comparative des phrases sans verbe en roumain et en français
Gabriela Bîlbîie (Author)
Einführung in die grammatische Beschreibung des Deutschen: Zweite, überarbeitete Auflage
Roland Schäfer (Author)
New perspectives on cohes
coherence: Implications fo
translation
Katrin Menzel, Ekaterina Lapshi
Kaltunski, Kerstin Kunz (Volume
Tone in Yongning Na: Lexical tones and morphotonology
Alexis Michaud (Author)
A grammar of Rapa Nui
Paulus Kieviet (Author)
Forthcoming: Attributive constructions in North-Eastern Neo-Aramaic
Ariel Gutman (Author)
Forthcoming: Empirical modelling of translation and interpreting
Silvia Hansen-Schirra, Oliver Čulo, Sascha Hofmann, Bernd Meyer (Volume Editor)
Forthcoming: Reflections on Bioliguistics
Cedric Boeckx (Author)
Diversity in African langua
Selected papers from the 4
Conference on African Ling
Doris L. Payne, Sara Pacchiarot
Bosire (Volume Editor)
Annotation, exploitation and evaluation of parallel corpora: TC31
Silvia Hansen-Schirra, Stella Neumann, Oliver Čulo (Volume Editor)
Forthcoming: Beiträge zur deutschen Grammatik
Tilmann N. Höhle (Author); Stefan Müller, Frank Richter, Marga Reis (Volume Editor)
The Ik language: Dictionary and grammar sketch
Terrill Schrock (Author)
A grammar of Papuan Malay
Angela Kluge (Author)
Forthcoming: Unity and diversity in grammaticalization scenarios
Walter Bisang, Andrej Malchukov (Volume Editor)
Tonal placement in Tashlh
intonation system accomm
adverse phonological envi
Timo B. Roettger (Author)
Forthcoming: Order and structure in syntax II: Subjecthood and argument structure
Laura R. Bailey, Michelle Sheehan (Volume Editor)
Forthcoming: Order and structure in syntax I: Word order and syntactic structure
Laura R. Bailey, Michelle Sheehan (Volume Editor)
Forthcoming: A aquisição da língua materna e não-materna: Questões gerais e dados do Português
Maria João Freitas, Ana Lúcia Santos (Volume Editor)
Forthcoming: Further investigations into the nature of phrasal compounding
Carola Trips, Jaklin Kornfilt (Volume Editor)
Dependencies in language: On the causal ontology of linguistic systems
N.J. Enfield (Volume Editor)
On looking into words (and
Structures, Relations, Ana
Claire Bowern, Laurence Horn,
Zanuttini (Volume Editor)

Workflow

- 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

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', *pare* '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-Szrejbrowska (2016) follows the idea that probing for phi-features is possible for T

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 (2018)

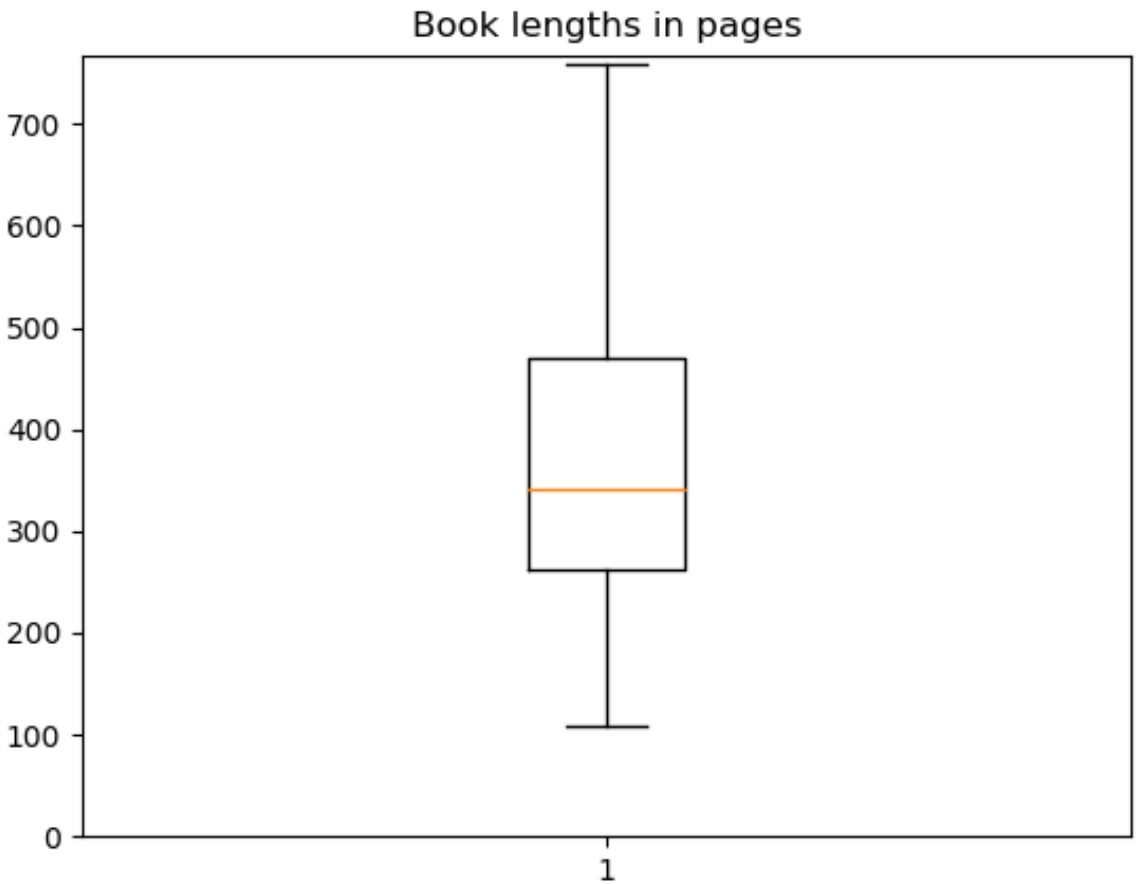
⟩ Westedt analysed a sample of comments on Paperhive for her BA thesis.

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

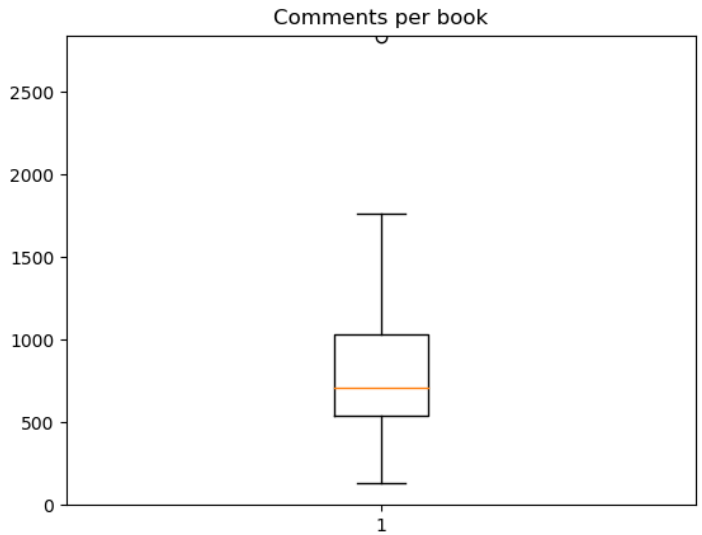
This study

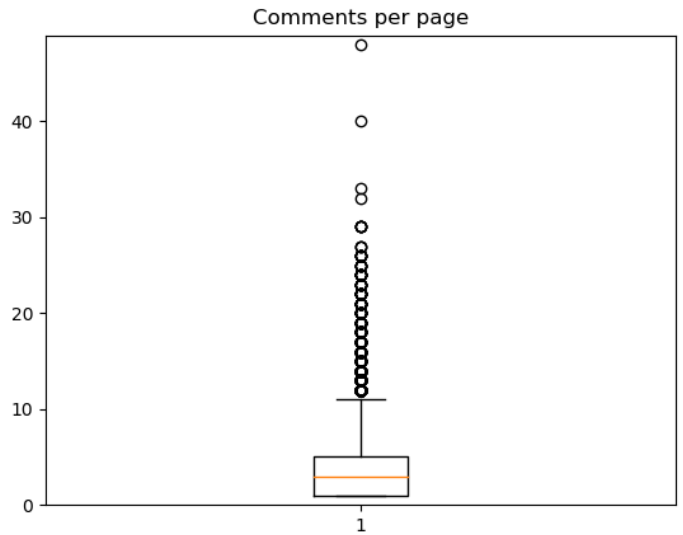
- 52 books from late 2016 to late 2018
- comments were harvested from Paperhive and put into a database
- 19 004 pages
- 43 370 comments

Book length



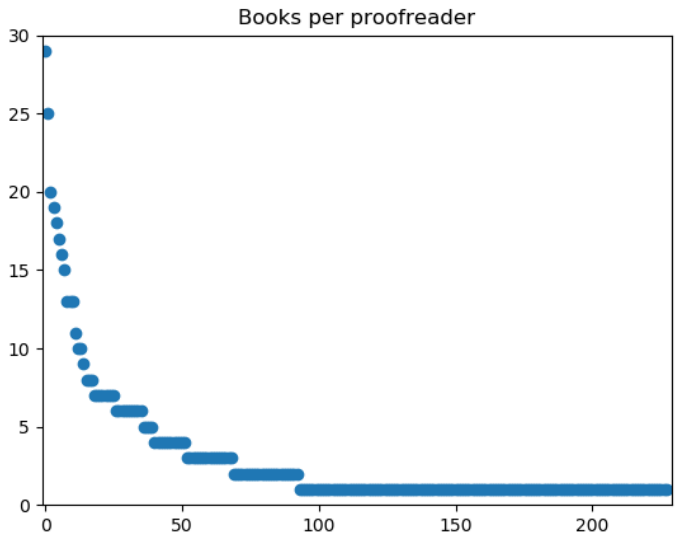
Comments

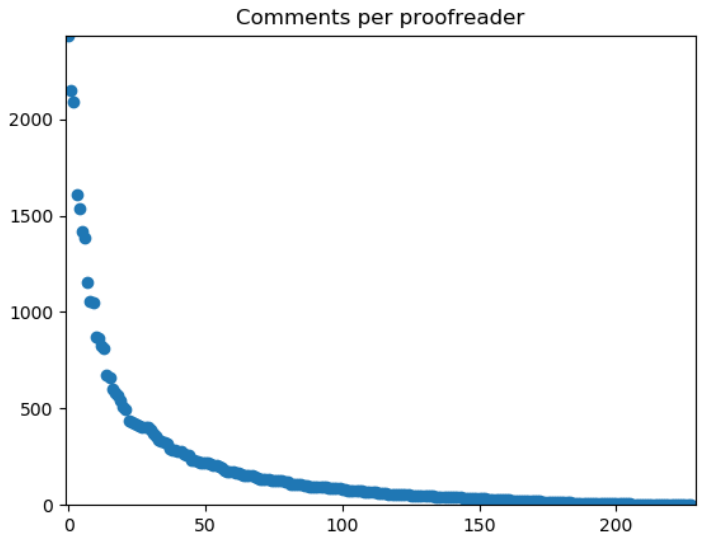




The highest number of comments on one page is found in Theory and description in African Linguistics on page 122 (48 comments).

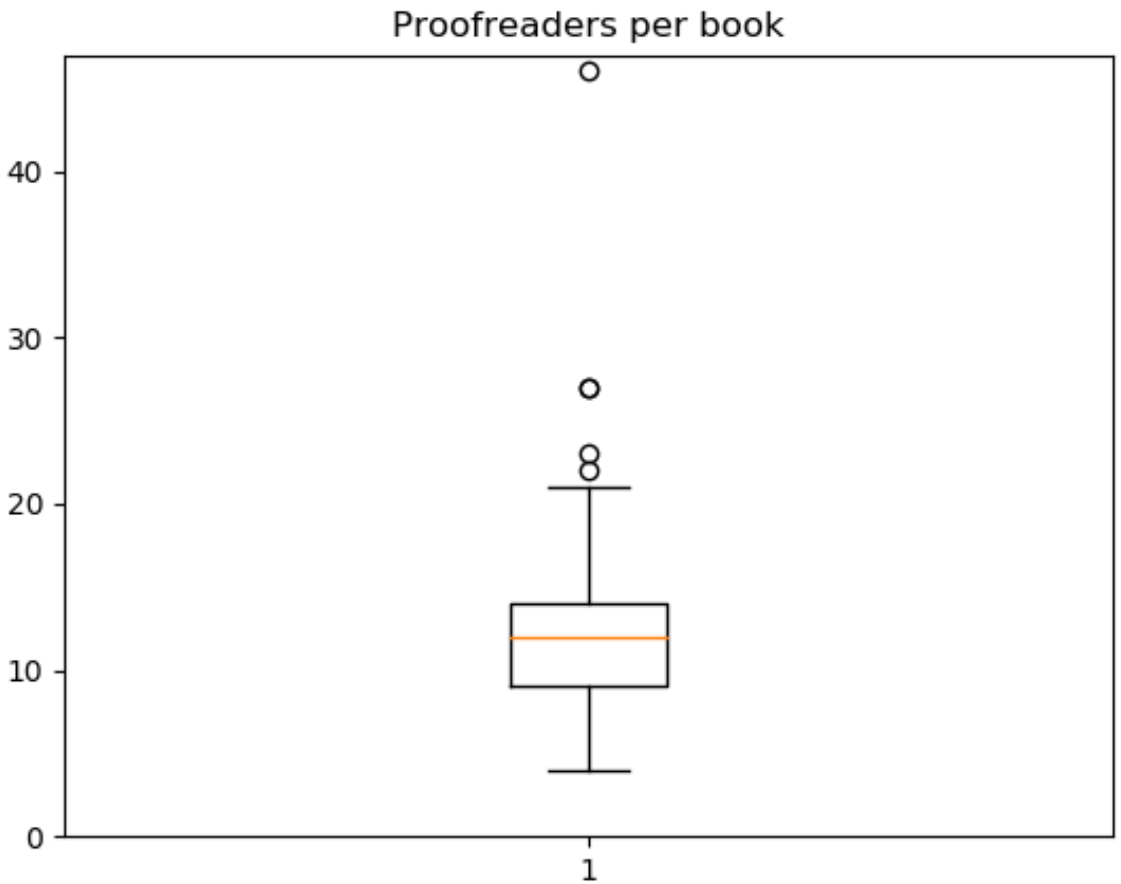
Productivity of proofreaders





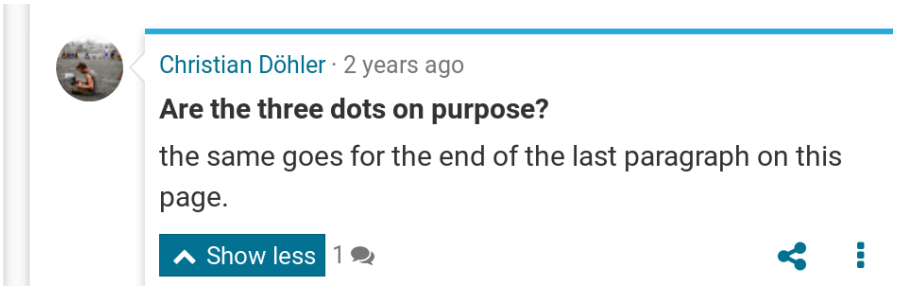
228 different accounts have participated in commenting.

Proofreaders per book



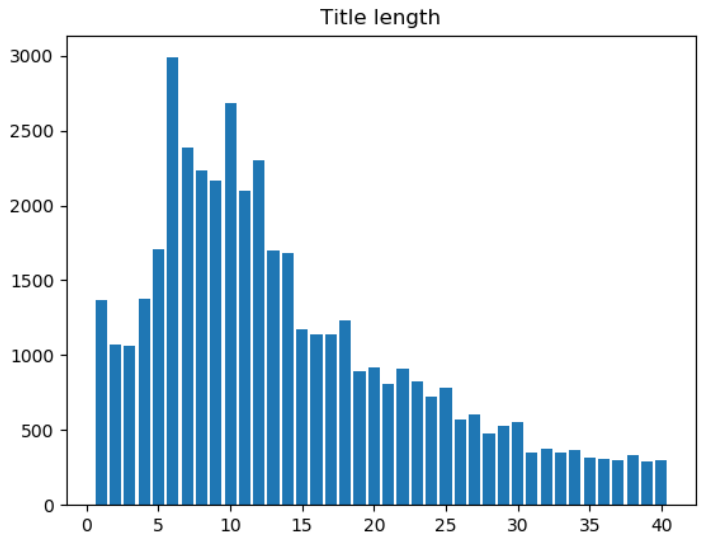
Text analysis

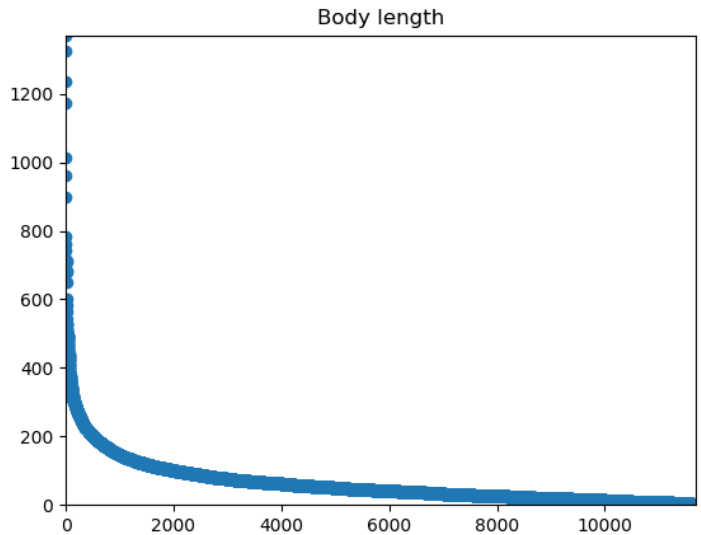
ourse about the Jewish Nerwa texts, where I rstood it as a recommendation to further my this...



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

Title length and body length





Hypotheses about proofreaders

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".**

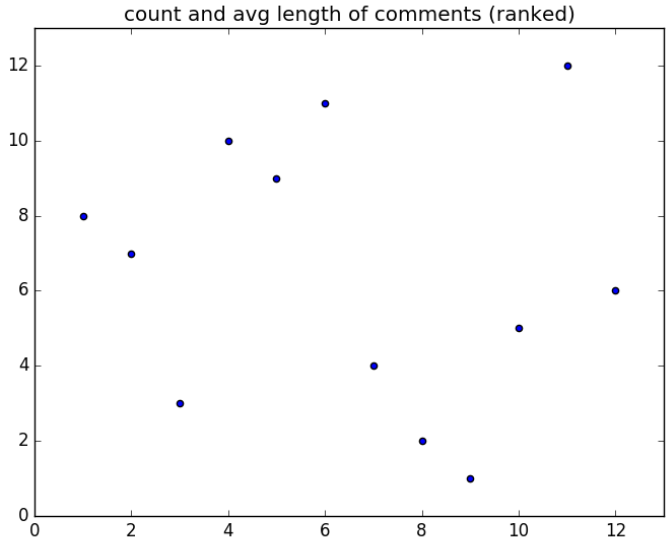
Hypothesis 1: proofreader types

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

Computation

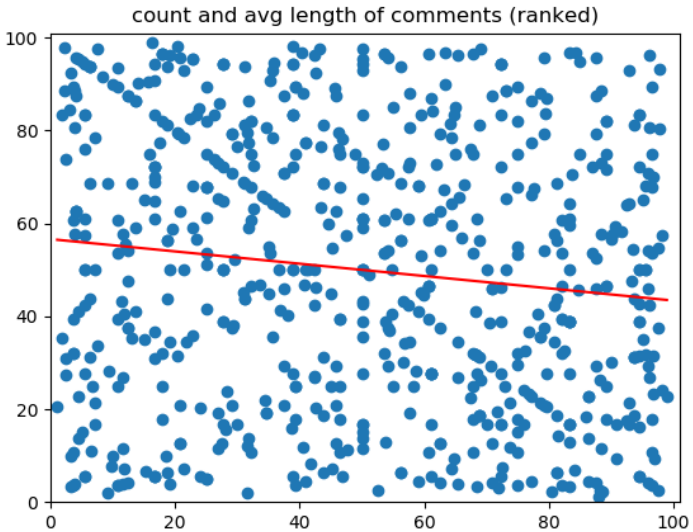
- For every book
  - rank all participating proofreaders by amount of comments
  - rank all participating proofreaders by average length of comments
  - plot the two against each other

Example of a plot for Hypothesis 1



- 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

Combination of all books



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

Result hypothesis #1

- Hypothesis #1 is confirmed
  - proofreaders with more comments have shorter comments
  - proofreaders with longer comments comment less

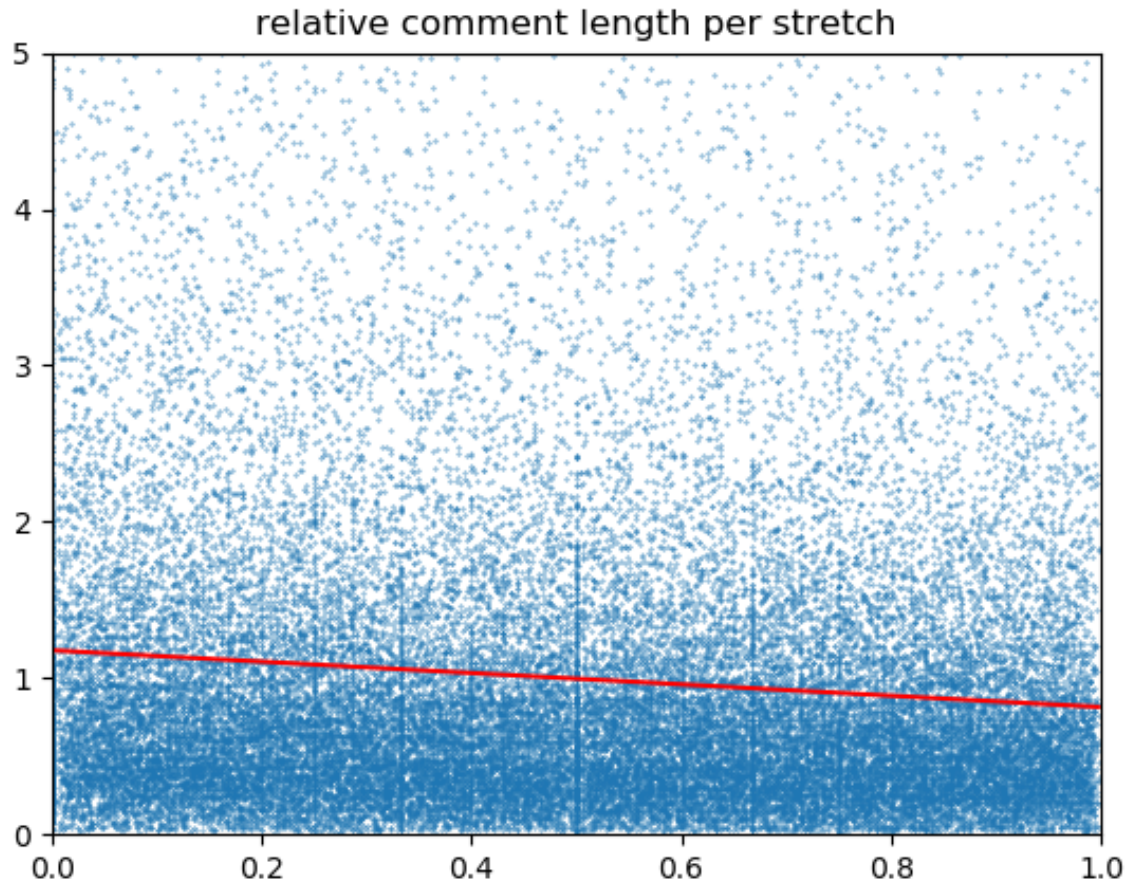
Hypothesis #2: proofreader fatigue

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”.

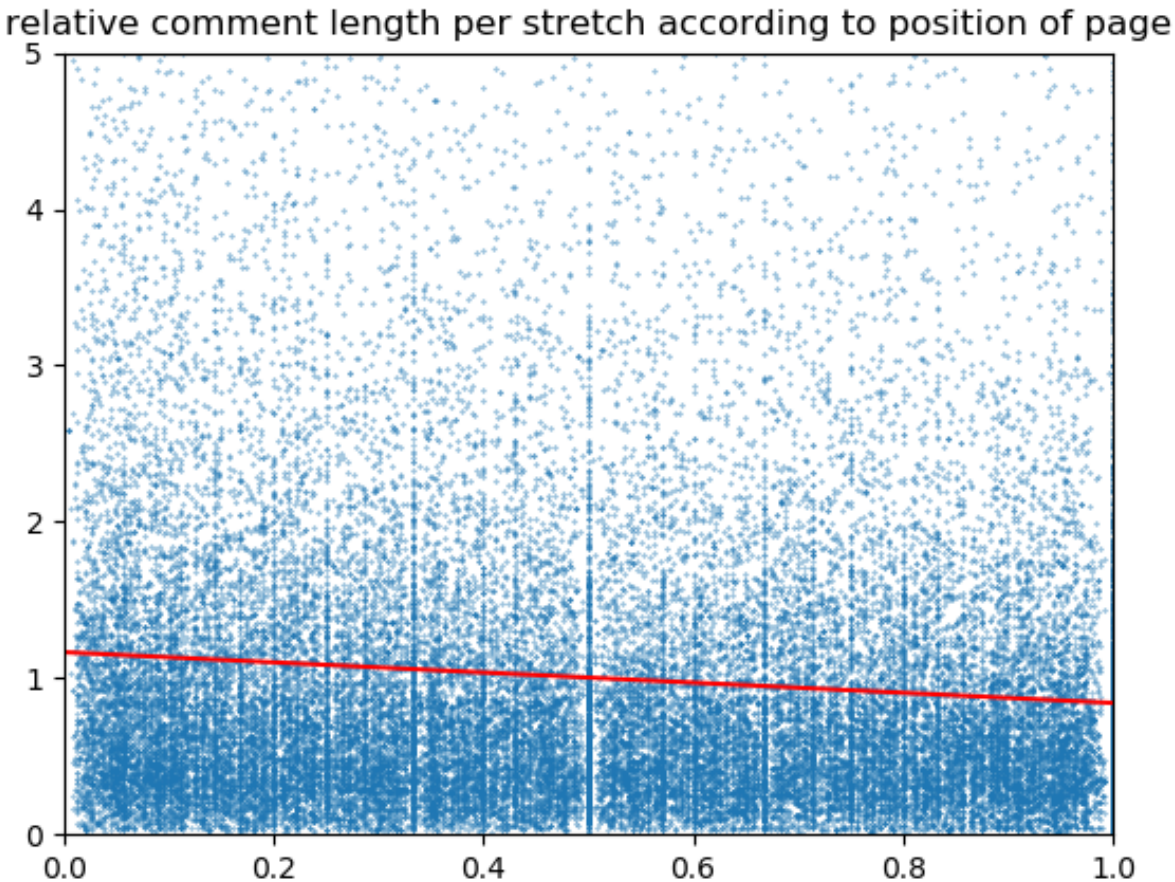
Computation for Hypothesis #2

- 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



Results for Hypothesis #2 “proofreader fatigue”

- 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

Discussion

- 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.

Do researchers take on different roles?

- 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.

Conclusions

- 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.

Questions

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

Thank you

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