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
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
<table>
<thead>
<tr>
<th>Books</th>
<th>Community proofreading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural causes of language: Frames, frames, and cultural transmission</td>
<td>Nj Efthymiou (Author)</td>
</tr>
<tr>
<td>Linguistic variation, identity construction and cognition</td>
<td>Nicola M. Ingram (Author)</td>
</tr>
<tr>
<td>Language strategies for the domain of colour</td>
<td>Jasmin Steiger (Author)</td>
</tr>
<tr>
<td>The empirical base of linguistics: Grammatical judgments and linguistic methodology</td>
<td>Ivan T. Spence (Author)</td>
</tr>
<tr>
<td>The evaluation of use grammar</td>
<td>Remco Taps (Author)</td>
</tr>
<tr>
<td>Tai in Yunnan: Na lexical tones and morfophonology</td>
<td>Alexa Richards (Author)</td>
</tr>
<tr>
<td>Annotation, exploitation and evaluation of parallel corpora: TC21</td>
<td>Silke Hartmann, Silke Hartmann, Silke Hartmann (Volume Editor)</td>
</tr>
<tr>
<td>Forthcoming: Order and structure in syntax: Subjecthood and argument structure</td>
<td>Laura B. Delhey, Michelle Sheehan (Volume Editor)</td>
</tr>
<tr>
<td>Forthcoming: Order and structure in syntax: Word order and syntactic structure</td>
<td>Laura B. Delhey, Michelle Sheehan (Volume Editor)</td>
</tr>
<tr>
<td>Forthcoming: A spuiable of language and meaning</td>
<td>Quedes garra de dos Portuguesas</td>
</tr>
<tr>
<td>A typology of marked-g-languages</td>
<td>Constance Kuipers (Author)</td>
</tr>
<tr>
<td>The Talking Heads experiment: Origins, effect and meanings</td>
<td>Luc L. Stroh (Author)</td>
</tr>
<tr>
<td>Grammaticization in the North: Russian, the languages of Scandinavia and Scandinavia versalita</td>
<td>Olaf Dahl (Author)</td>
</tr>
<tr>
<td>A grammar of Pito Sami</td>
<td>Jorun Riede (Author)</td>
</tr>
<tr>
<td>Proustian detail in Neoplatonic</td>
<td>Francesco Cappellari (Author)</td>
</tr>
</tbody>
</table>
 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/*weszły* 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 dom.
    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
Westedt Analysed a sample of comments on Paperhive for her BA thesis.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style</td>
<td>21.00</td>
</tr>
<tr>
<td>Lexical choice</td>
<td>20.73</td>
</tr>
<tr>
<td>Punctuation</td>
<td>11.81</td>
</tr>
<tr>
<td>Grammar</td>
<td>11.55</td>
</tr>
<tr>
<td>References</td>
<td>9.71</td>
</tr>
<tr>
<td>Syntax</td>
<td>7.80</td>
</tr>
<tr>
<td>Spelling</td>
<td>7.30</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td><strong>6.56</strong></td>
</tr>
<tr>
<td>Miscellanea</td>
<td>3.41</td>
</tr>
</tbody>
</table>
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
› 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

Descriptive statistics

228 different accounts have participated in commenting.
Proofreaders per book
A PaperHive comment has a succinct title (<40 characters)

optional body, with more elaborate information
Title length and body length

Descriptive statistics

Title length

Body length

[Graphs showing title and body length distributions]
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
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
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: 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

relative comment length per stretch

Hypothesis evaluation | Proofreader fatigue
Plot for Hypothesis #2 based on page position

relative comment length per stretch according to position of page
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.
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, such as proofreader typology and proofreader fatigue. Flow back and forth between the group of authors and the group of proofreaders creates a healthy ecosystem where 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?
Gold proofreaders

- Andreas Hölzl (view profile) 34/99
- Jeroen van de Weijer (view profile) 31/99
- Eitan Grossman (view profile) 29/99
- Jean Nitzke 25/88
- 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