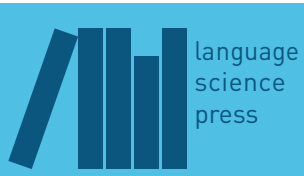
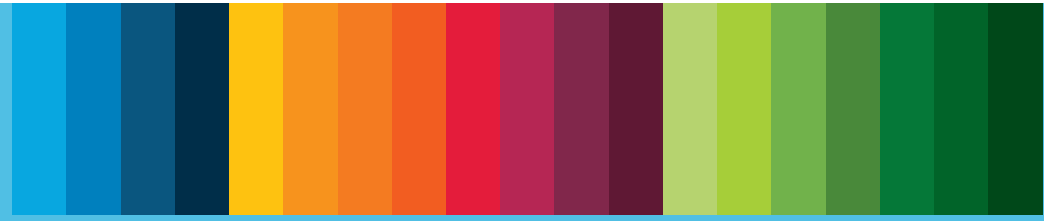




Technical, legal, and social aspects of Open Science in linguistics

Sebastian Nordhoff

CNRS-Villejuif, 2021-09-09

Sebastian Nordhoff

- PhD 2009 *A grammar of Upcountry Sri Lanka Malay*
- 3 edited volumes, about 30 research articles
- 2009-2012 Glottolog.org at the Max Planck Institute for Evolutionary Anthropology
- advocate for Open Source, Open Access, Open Data, Open Everything
- since 2014 coordinator for Language Science Press
- involved in the publication of 150+ books since 2014

Language Science Press

- scholar-owned, community-based publisher
- Open Access
- free to read
- free to publish
- 29 series
- 30 books a year (monographs and edited volumes)

Language Science Press

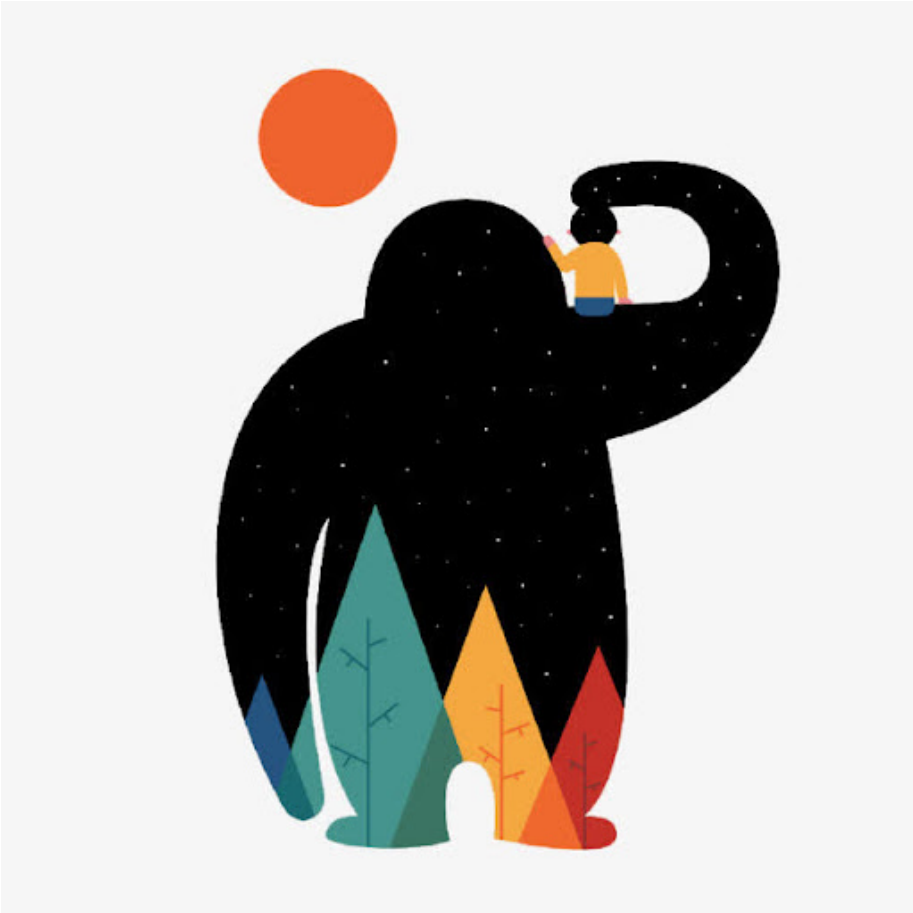
Outline of this talk

1. Open Science
2. Workflow of a scientific manuscript
3. How can each step in the workflow be made more open?

Goals of open science

1. replicable/reliable
2. inclusive
3. participative
4. collaborative
5. transparent
6. reusable

⟩ → allow other people to stand on your shoulders



Scientific workflow

1. **gathering of data**
   - ⟩ eg recordings, questionnaires
2. **transformation**
   - ⟩ cut, crop, cleanse
3. **enrichment**
   - ⟩ link to other data, analyze
4. **publication**
   - ⟩ make available

Steps

1. pre-submission
2. submission
3. review
4. revision
5. proofreading
6. publication

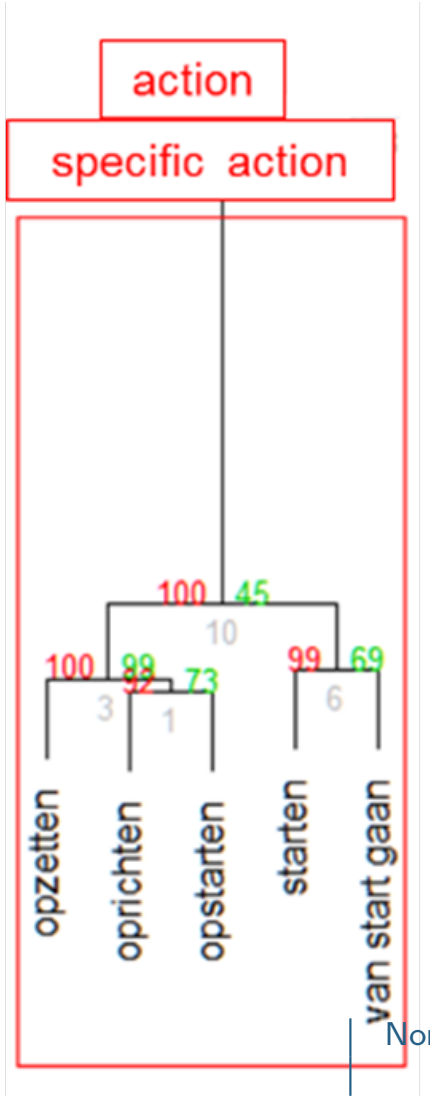
7. reading
8. iteration

⟩ prestige

⟩ long-term preservation

Pre-submission: object types

- Scientific research uses different kinds of entities
  - data
  - scripts/processes/methods
  - texts
- Do **keep them separate**
- allow for **modular changes**
- the image to the right was created with R, but the data was hard coded in the file.
- there are 100 images like that; for each one, the script was changed
- → impossible to recreate the 100 files with legible colours.



Pre-submission: formats

- Documents come in different formats (pdf, jpg, mov)
- For your work to be accessible to others, you should use fileformats which are
  - standardized
    - a skilled and motivated person should be able to write an application to read and write the format
    - otherwise, your file risks ending up in a data graveyard
  - open
    - royalty-free
  - suited for tasks
    - eg, do not use jpg for texts
- `https://en.wikipedia.org/wiki/List_of_open_formats`

Sample table

- what an easy table looks like to humans

	Nouns	Verbs	Adjectives
Group A	99	149	122
Group B	152	141	131

- what an easy table looks like to computers

Sample table

⟩ what an easy table looks like to humans

	Nouns	Verbs	Adjectives
Group A	99	149	122
Group B	152	141	131

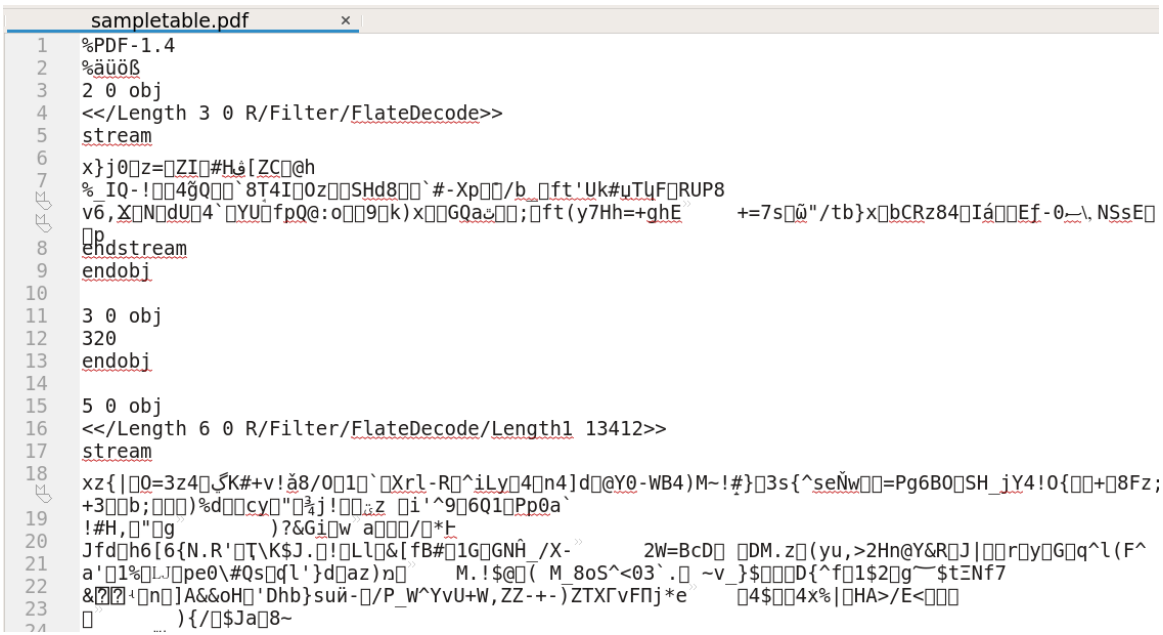
⟩ what an easy table looks like to computers **png**:

Sample table

⟩ what an easy table looks like to humans

	Nouns	Verbs	Adjectives
Group A	99	149	122
Group B	152	141	131

⟩ what an easy table looks like to computers **pdf**:



Sample table

⟩ what an easy table looks like to humans

	Nouns	Verbs	Adjectives
Group A	99	149	122
Group B	152	141	131

⟩ what an easy table looks like to computers **LibreOffice XML**:

```
−<office:document-content office:version="1.2">
   <office:scripts/>
  +<office:font-face-decls></office:font-face-decls>
  +<office:automatic-styles></office:automatic-styles>
  −<office:body>
    −<office:text>
      +<text:sequence-decls></text:sequence-decls>
      −<table:table table:name="Table1" table:style-name="Table1">
          <table:table-column table:style-name="Table1.A" table:number-columns-repeated="4"/>
        −<table:table-row>
          −<table:table-cell table:style-name="Table1.A1" office:value-type="string">
              <text:p text:style-name="Table_20_Contents"/>
            </table:table-cell>
          −<table:table-cell table:style-name="Table1.A1" office:value-type="string">
              <text:p text:style-name="P1">Nouns</text:p>
            </table:table-cell>
          −<table:table-cell table:style-name="Table1.A1" office:value-type="string">
              <text:p text:style-name="P1">Verbs</text:p>
            </table:table-cell>
          +<table:table-cell table:style-name="Table1.D1" office:value-type="string"></table:table-cell>
          </table:table-row>
        −<table:table-row>
          −<table:table-cell table:style-name="Table1.A2" office:value-type="string">
              <text:p text:style-name="P1">Group A</text:p>
            </table:table-cell>
          −<table:table-cell table:style-name="Table1.A2" office:value-type="string">
              <text:p text:style-name="P1">99</text:p>
            </table:table-cell>
          −<table:table-cell table:style-name="Table1.A2" office:value-type="string">
              <text:p text:style-name="P1">149</text:p>
            </table:table-cell>
          −<table:table-cell table:style-name="Table1.D2" office:value-type="string">
              <text:p text:style-name="P1">122</text:p>
            </table:table-cell>
          </table:table-row>
        −<table:table-row>
```

Sample table

- what an easy table looks like to humans

	Nouns	Verbs	Adjectives
Group A	99	149	122
Group B	152	141	131

- what an easy table looks like to computers **csv**:

Documents

sampletable.csv

```
1   Nouns,Verbs,Adjectives
2   Group A,99,149,122
3   Group B,152,141,131
4
```

Dissemination

- For your work to be useful for others, those "others" must
  - know that it exists (**find** it)
  - be able to **access** it
- preprint servers
  - lingbuzz
  - semanticsarchive
  - Rutgers Optimality Archive
- repositories
- publishers

Repositories

- Repositories collect documents and research data
  - **discipline-specific** or **general purpose**
  - **preprint** or **postprint**
  - **text** or **data**

Replication Data for: Russian verbal borrowings in Udmurt

May 14, 2019

Arkhangelskiy, Timofey, 2019, "Replication Data for: Russian verbal borrowings in Udmurt", https://doi.org/10.18710/5N34CG, DataverseNO, V1

This is the dataset used in a study of Russian verbal loans in Udmurt. The files contain lists of Russian verbs found in the Udmurt social media corpus (http://udmurt.web-corpora.net/index_en.html), manually annotated for several features such as aspect or frequencies in differen...

Replication Data for: Accusative of Negation in 'Borderland' Polish

Mar 8, 2019

Fellerer, Jan, 2019, "Replication Data for: Accusative of Negation in 'Borderland' Polish", https://doi.org/10.18710/CYPRAY, DataverseNO, V1

These are the data for a journal article on 'Accusative of Negation in 'Borderland' Polish'. The abstract of the article is below. The data consist of the annotated list of tokens of accusative vs. genitive of negation (=GenNeg.txt), excerpted manually from relevant sources docum...

Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-

Feb 6, 2019

Rainsford, Thomas, 2019, "Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-", https://doi.org/10.18710/X5ZFXZ, DataverseNO, V1

This dataset contains the raw data and the R scripts necessary to replicate all tables and figures in the cited publication. The raw data consists of manually-annotated plain-text concordances containing instances of five pairs of Old French spatial prepositions (ens/dedans, hors...

Replication Data for: Seeing from without, seeing from within: aspectual differences between Spanish and Russian

Nov 27, 2018

Janda, Laura A; Fábregas, Antonio, 2018, "Replication Data for: Seeing from without, seeing from within: aspectual differences between Spanish and Russian", https://doi.org/10.18710/WR4Y0Q, DataverseNO, V1, UNF:6:v5Lkz2Vq1VjqBSIUTbLvrA== [fileUNF]

This is the data that serves as the basis for an article comparing the grammatical category of aspect in Spanish and Russian. Here is the abstract of the article: Linguistic categories such as aspect are not identical across languages, and cross-linguistic differences can reveal...

Welcome to LingBuzz, an article archive and a community space for Linguistics. You are highly encouraged to upload your articles - old and new, published or not. LingBuzz is maintained by Michal Starke and hosted by Tromso. [more about LingBuzz]

The Buzz

[publish a paper]

search | phonology go

Mayer, Connor Daland, Robert	**freshly changed**	[pdf]	A method for projecting features from observed sets of phonological classes
Seveleu-Dubrovnik, Maxime	**freshly changed**	[pdf]	Inchoative-causative alternation in Persian
Al-Khalaf, Eman	**freshly changed**	[pdf]	Floating Quantifiers are Autonomous Phrases: A Movement Account
Xiang, Yimei	**freshly changed**	[pdf]	Function alternations of the Mandarin particle dou: Distributor, free choice licensor, and 'even'
Adger, David	**new**	[pdf]	Linguistic Representations: a note on terminology vs. ontology

Top Recent Downloads

1. Cheshire - Linguistically Dating and Locating Manuscript MS408
2. Cheshire - Linguistic Missing Links.
3. Pesetsky - Exfoliation: towards a derivational theory of clause size

Papers from as of 2019 05 28:
[semanticsarchive.net]

2019 05 26 Frana, Ilaria and Kyle Rawlins
Attitudes in Discourse: Italian Polar Questions and the particle "mica"
2019 05 23 Frana, Ilaria
Concealed Questions (SEMCOM)
2019 05 20 Greenebrg Yael
Even and Only: Arguing for parallels in scalarity and in constructing alternatives
2019 05 20 Greenberg, Yael and Lavi Wolf
Intensified response particles: The case of Hebrew 'legamrey'
2019 05 18 Daniel Büring
Topless and Salient — Convertibles in the Theory of Focus
2019 05 17 Šimík, Radek
Inherent vs. accidental uniqueness in bare and demonstrative nominals
2019 05 17 Wagner, Michael
Prosodic Focus
2019 05 08 Marty, Paul and Romoli, Jacopo
Presuppositions, implicatures, and contextual equivalence
2019 05 06 Capone Alessandro
Pragmatics and Philosophy. Connections and ramifications
2019 05 03 Yanovich, Igor
Epistemic Modality [survey article]
2019 05 02 Kuhn, Jeremy
Positive uses of NPIs and logical duality [squib]

Rutgers Optimality Archive

Rutgers **O**ptimality **A**rchive

[Author Login]

The Rutgers Optimality Archive is a distribution point for research in Optimality Theory an
Posting in ROA is open to all who wish to disseminate their work in OT and related theorie

Policies
Optimal List
About ROA

SEARCH Text Advanced Search

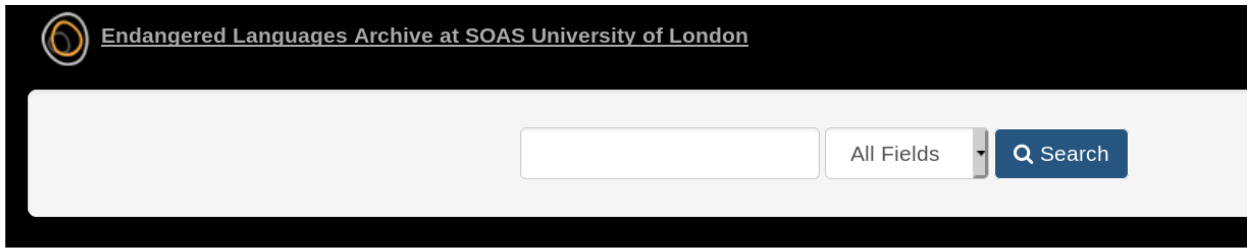
BROWSE List All Postings
View by ROA number: Search

SUBMIT Create a new submission
View your previous submissions

Recent Submissions

Click on ROA# to download. Click on Title to view abstract.

ROA #	Authors	Title
ROA 1356	Canaan Breiss, Bruce Hayes	Phonological markedness effects in sentence formation
ROA 1355	Will Bennett, Natalie DelBusso	ABC(D) Book
ROA 1354	Gulab Chand , Somdev Kar	Sonority and Reduplication in Hadoti
ROA 1353	Veltzer Doron	Mechanical Turkish



Search: Africa

Level

Deposit (88)

Funding body

ELDP (65)
National Science Foundation (6)
AHRC (2)
GbS (1)
National Endowment for the Humanities (1)
SOAS University of London (1)

Status

Collection online (82)
Forthcoming (6)

Country

Showing **1** - **10** of **88** for search: **'Africa'**
,query time: 0.05s

Deposit

A Documentation of Gurene Folk Tales, Riddles, Songs, Palace Genres and other Oral Genres in Bolga

Depositor: Samuel Atintono

The documentation is largely a collection of audio and video recordings of endangered traditional folktales and riddles narrated by expert narrators in Bolgatanga (Bolga) in the Upper East Region of Ghana between 2010 and 2011.

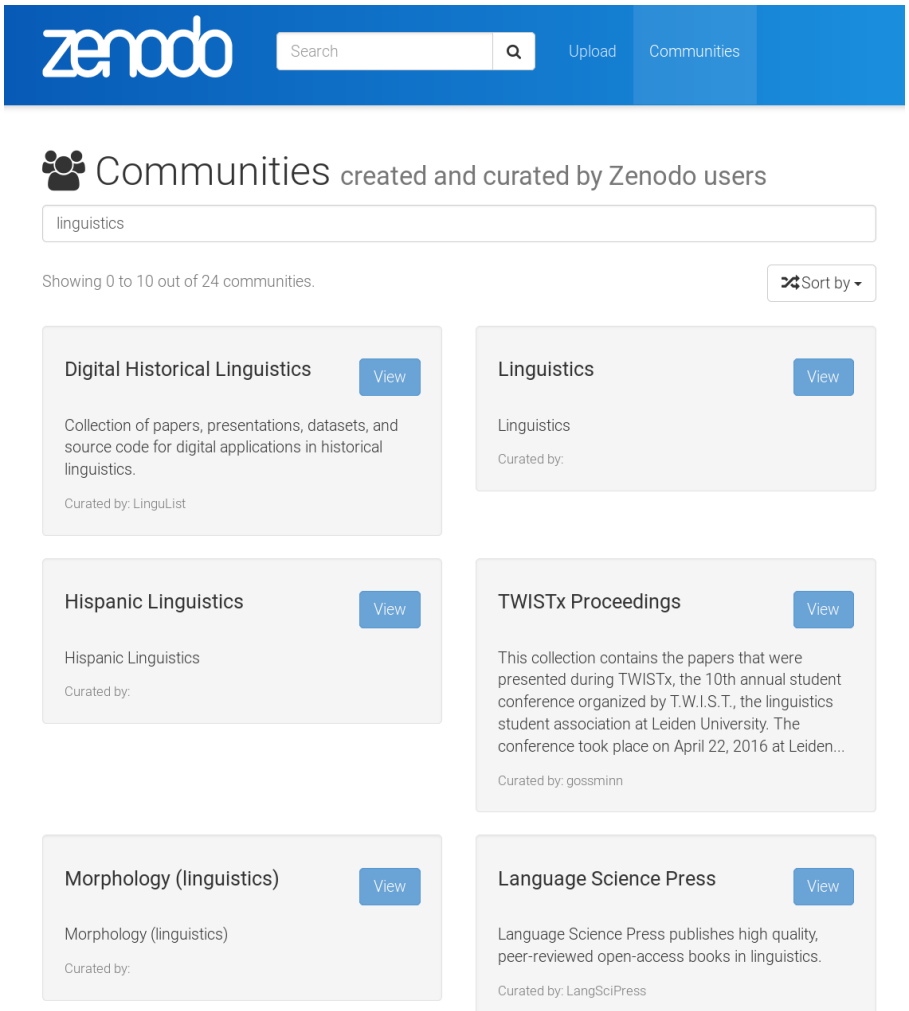
Keywords: Gurene

Deposit

A Documention of Tabaq, a Hill Nubian language of the Sudan, in its sociolinguistic context

Depositor: Gerrit J. Dimmendaal, Birgit Hellwig

Zenodo


zenodo
Search
Upload
Communities
Communities created and curated by Zenodo users
linguistics
Showing 0 to 10 out of 24 communities.
Sort by
Digital Historical Linguistics
View
Collection of papers, presentations, datasets, and source code for digital applications in historical linguistics.
Curated by: LinguList
Linguistics
View
Linguistics
Curated by:
Hispanic Linguistics
View
Hispanic Linguistics
Curated by:
TWISTx Proceedings
View
This collection contains the papers that were presented during TWISTx, the 10th annual student conference organized by T.W.I.S.T., the linguistics student association at Leiden University. The conference took place on April 22, 2016 at Leiden...
Curated by: gossminn
Morphology (linguistics)
View
Morphology (linguistics)
Curated by:
Language Science Press
View
Language Science Press publishes high quality, peer-reviewed open-access books in linguistics.
Curated by: LangSciPress

Academia and Researchgate

- Academia.edu and researchgate are NOT repositories
- both make money from restricting access

A note on scooping

- Scooping means that someone “steals” your research while you are still finalising it
- that fear is by and large unwarranted
- you can count yourself happy if anybody IS actually interested in your data!
- most research actually struggles a lot more with a lack of interest to outsiders rather than scooping
- preprint servers allow you to register your data and establish primacy

Intellectual Property Rights

- “Copyright” in the Anglo-Saxon countries
- “Urheberrecht” in continental Europe
- These are different!
- Once you create somehting, you can decide who can use it and how
- Choose wisely and be explicit!

License

- Usage rights can be differentiated as follows
  - geographical restriction (e.g. *only for Germany*)
  - type restrictions (e.g. *only for print*)
  - exclusive (no-one else has the right) vs. non-exclusive (other people may also get the rights)
  - copyright transfer agreement (Anglo-Saxon culture)
  - total buy-out

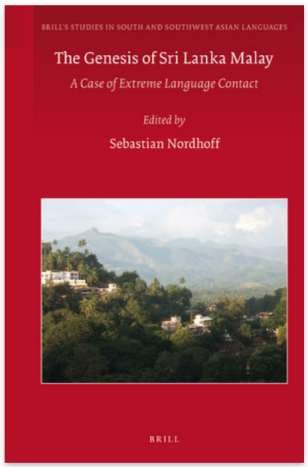
Copyright transfer agreement

TRANSFER OF COPYRIGHT

12. The Editor hereby assigns to the Publisher the full copyright in the Work, which assignment the Publisher hereby accepts. Consequently, the Publisher shall have the exclusive right throughout the world to publish and sell the Work in all languages, in whole or in part, including, without limitation, any abridgement and substantial part thereof, in book form and in any other form including, without limitation, mechanical, digital, electronic and visual reproduction, electronic storage and retrieval systems, including internet and intranet delivery and all other forms of electronic publication now known or hereinafter invented. The assignment of copyright shall include the name of the Work and any additions and alterations in the event of a revision.

Contracts

- Publisher contracts restrict everybody, including yourself
- once you sign away your copyright, you have no longer the right to use your own material
- to reuse tables or graphics in subsequent works of yours, you must first ask the new rights holder for permission
- chasing rights is incredibly annoying
- publishers will put your content behind a paywall, meaning that it can actually be more difficult to access once it is officially published than before
- publisher vary as to whether and when they allow books to be hosted in repositories



The Genesis of Sri Lanka Malay

A Case of Extreme Language Contact

Series:
Brill's Studies in South and Southwest Asian Languages, Volume: 3

Editor: Sebastian Nordhoff

In *The Genesis of Sri Lanka Malay: A Case of Extreme Language Contact*, the synchrony and diachrony of Sri Lanka Malay are investigated from a variety of angles: Experts on South Asia, South East Asia, Creole Studies, Areal Linguistics, Typolog
See More

Publication Date: 29 November 2012
ISBN: 978-90-04-24225-8
DOI: https://doi.org/10.1163/9789004242258

Open Access colour codes

- green
  - Normal copyrighted publication with a publisher, but copy is archived in an institutional repository
- gold
  - publication is made openly available against a fee (Article Processing Charge, Book Processing Charge)
- diamond
  - like Gold OA, but without a fee
- (black)
  - a copyrighted publication is available via pirate sites/shadow libraries like SciHub or LibGen
- (bronze)
  - fake open access, not respecting the Berlin Declaration

Publisher selection

⟩ Fair Open Access Principles (`https://www.fairopenaccess.org/the-fair-open-access-principles`)

1. The journal has a **transparent ownership structure**, and is **controlled by** and responsive to the **scholarly community**.
2. Authors of articles in the journal **retain copyright**.
3. All articles are published open access and an **explicit open access licence** is used.
4. **Submission and publication is not conditional** in any way on **the payment of a fee** from the author or their employing institution, or on membership of an institution or society.
5. Any **fees** paid on behalf of the journal to publishers **are low, transparent, and in proportion** to the work carried out.

Publisher selection

- journals: `oaling.wordpress.com` has a list of diamond (no fee) OA journals
- books: not so many diamond options, besides LangSci Press

Submission

- How could a more open and more inclusive submission process look like?
  - no formal requirements
  - you do not have to follow all house rules for the initial submission. This is only necessary once the manuscript is approved.

Review

- How could a more open reviewing process look like?
- Open Peer Review
  - prepublication/postpublication
  - invited/self-selected
  - influences decision to accept Y/N
  - review text openly available/hidden
- Code/Workflow review
  - If there are quantitative arguments in the paper, are data and computer code available so that interested readers can try to replicate and evaluate the findings?
  - H*ow to Share Data and Code when Submitting Papers to a Journal: Practical Questions*
    `https://calc.hypotheses.org/2782`

Revision

- So your manuscript is accepted with major revisions. What next?
  - **content revision**: fine-tune argumentation
  - **formal revision**: make sure all guidelines are met and tables and figures look good
  - Overleaf.com allows for a separation of labour:
    - author does the content revision
    - student assistants take care of the menial tasks
    - typesetters (yours truly) take care of tables, diagrams, and complicated issues.

Proofreading

- How could a more open proofreading/copyediting process look like?
- Get rid of publisherese and get in touch with “real” readers from different demographics
- Community Proofreading
- Each of the LangSci books is read by 10-20 proofreaders (1-2 chapters per person)
- All of them have different background
- Heterogeneous comments, but should be able to discover problems for different groups of readers.

Publication/Dissemination

- So the revised, typeset pdf is there. How would a more open approach to dissemination look like?
- no fees.
- easy download
- no vendor lock-in “reading tools” in the browser
- next to the pdf, the raw input material (numerical figures, graphics, computer code) is made available
- for people who prefer paper, print-on-demand is offered.
- Fair Open Access Principles: keep your copyright!

Reading

- How would a more open reading experience look like?
- no fees for the reader
- format neutral
  - next to pdf, also other formats like HTML or epub are proposed
    - LangSci does not offer this as of today

Iteration

- After publication, the readership will interact with the work. How can this be made more open?
- shorter iteration cycles
  - make it easy to have subsequent editions to incorporate feedback
- Post Publication Peer Review
  - allow readers to easily leave comments about certain passages
  - “collaborative reading” with PaperHive
- “forkability”
  - allow others to “rebuild” your book based on the basic materials with amendments
- use the Creative Commons Attribution license (CC-BY)

Prestige

- some venues enjoy a lot of prestige, others don't, but the reasons for this are often historic and very often completely opaque
- open download figures can provide some transparency here
- but note that commodification/quantification is very often a bad proxy to judge scientific quality

Long term preservation

- Special mechanisms have been devised to make sure that cease of operation of publishers does not entail that the works published by them are lost
- CLOCKSS (*Lots of Copies Keep Stuff Safe*) for legacy publishers
- trigger events
  - eg bankruptcy
- for open publications: the trigger event is simply the initial publication. No need for contrived processes like CLOCKSS

Wrap-up: recommendations

- separate data, code, and text
- publish all your data in appropriate repositories
- use preprint servers

- use the CC-0 license for data and the CC-BY license for text
- do not sign away your copyright

- care for your discipline, go for scholar-led community-owned publishers

Thank you


Check List 1738
The Pierpont Morgan Library
Nebrissensis, las introducciones latinas.
(Salamanca, primer grupo gótico, ca. 1486.)
Haebler 462. Proctor 9560.
Sonst von keinem Bibliographen erwähnt.
Nur noch 1 Exemplar in Brit.Mus. bekannt.