Social Learning of Knowledge Representations in Wikipedia

Peeter Tinits[1], Stefan Hartmann[2]
1-Tallinn University; 2-University of Bamberg

Introduction

- When we present knowledge about the world, it is useful to represent similar things in similar ways. This helps an experienced reader (conventions) and may grasp something about reality (iconicity).
- This can be a difficult task: humans are notoriously bad at maintaining clean categories of information[1].
- In this project, we study the degree of information standardization in Wikipedia.

- Are Wikipedias becoming more structured?
- Is that structure becoming conventional?
- What mechanisms support it?

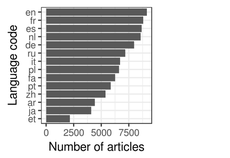
An exploratory corpus-based social learning study.

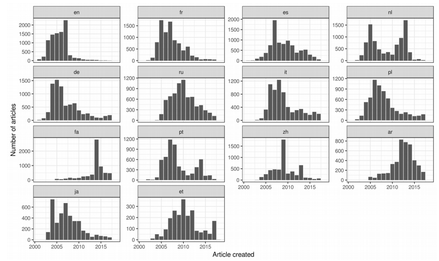
Methods

Object
- Wikipedia - collaborative encyclopedia.
- ~50M users, 200k active writers.
- Ideally, one phenomenon – one article.
- Information distributed into subsections based on aspects of the phenomenon.

Data
- Corpus of Wikipedia subsection headers
- Article histories on 9189 bio. species, described in at least 20 Wikipedias.
- 13 most active Wikipedias + Estonian
- Presence or absence of headers in time

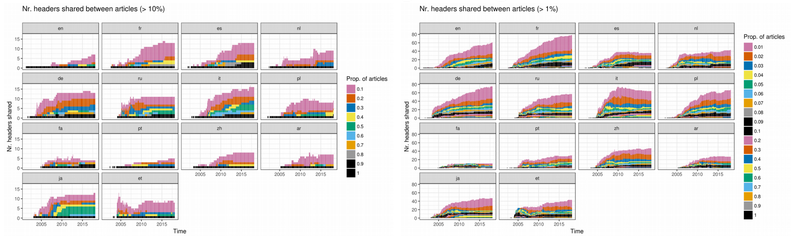
lang	year	ARTICLE	HEADER	present
en	2005	Giant pangolin	behavior	1
en	2005	Giant pangolin	physical description	1
en	2005	Giant pangolin	sources	1
en	2007	Giant pitta	introsection	1
en	2007	Giant pitta	source	1



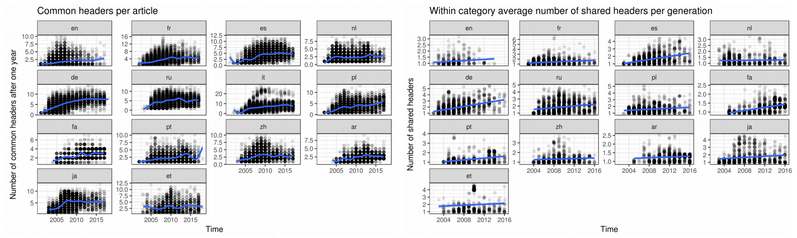


Increase in structure

Standardization of headers:
- Over time, the number of headers shared between the articles in our set is growing.



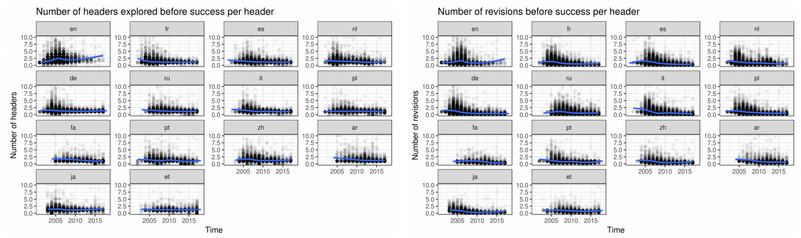
"Generations" of articles from each year
- Popular headers (> 1% articles) after 365 days
- Number of headers per article
- Average number of headers shared within category



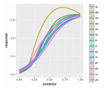
Are these conventions?
- We can check if particular headers in new articles are based on earlier prevalence in the set.
- Different sets can be made: e.g. accumulated frequencies, frequencies weighted by pageviews, frequencies boosted by prestige, additions the "generation" before, etc.
- This can be done within categories (here: *mammals, birds, snakes etc)* or across the dataset. See more in top right.

Search space

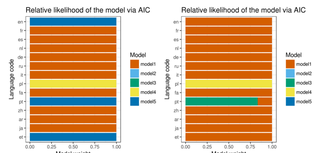
Are newer articles quicker to find the "right" headers?



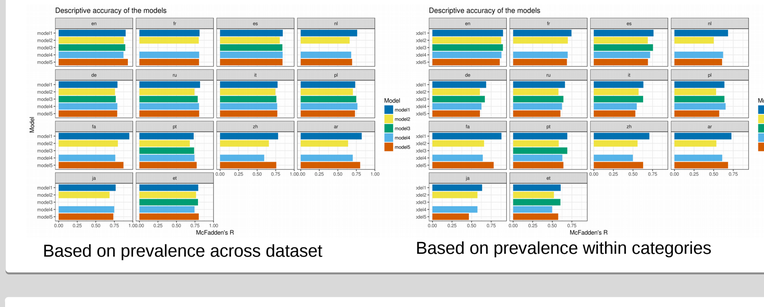
Social learning models



We contrasted plausible learning strategies[2]:
1) Simple replication
2) Weighted by popularity
3) Weighted by prestige
4) Recent additions
5) Average



We fit logistic regression models with different learning sets, and used AIC for relative likelihood of the models.



Discussion

- Wikipedias are getting more structured, however with notable language differences.
- Digital data offer rich resources to study social learning mechanisms[3], however often they may be hard to distinguish.
- Here, simple replication seems fit best, but others are close by.

Future aims:
- Better ways to distinguish learning trajectories when they make similar predictions.
- Explanations for differences between Wikipedias: Community size? Cultural differences?
- A closer look at translational equivalence of articles.

References

1. Latour, B. (1993). *We have never been modern*. Harvard University Press
2. Kendal, R. L., Boogert N.J., Rendell, L., Laland, K.N., Webster, M., Jones P.L. (In press). Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*.
3. Acerbi, A. (2016). A Cultural Evolution Approach to Digital Media. *Front. Hum. Neurosci.*

TALLINN UNIVERSITY

UNIVERSITÄT BAMBERG

Correspondence address: peeter.tinits@gmail.com