Social learning of knowledge representations in Wikipedia

PEETER TINITS*[1], STEFAN HARTMANN[2]

* Corresponding author: peeter.tinits@gmail.com

[1] *Tallinn University*

[2] *University of Bamberg*

Recorded history knows many attempts to gather and systematize knowledge to represent it in an efficient and accessible way, ranging from historical chronicles through enlightenment encyclopedias to online discussion boards. Typically, this efficiency is found in describing similar phenomena in similar terms: standardization of representations follow the expectations of an experienced reader and gradually better and more effective ways to represent the phenomena may be found.

We investigate the degree of standardization of knowledge representations in the edit history of Wikipedia and the potential role of social learning in supporting it. Particularly, we consider the aspects covered in Wikipedia articles as represented by their section headers. For example, the article on platypus has sections on “Description”, “Evolution”, “Conservation”, as well as “Venom” and “Electrolocation”. Other animals are likely to include also the first three, while “Venom” can be expected only for venomous animals.

We analyse the prevalence of different headers in a custom multilingual corpus in Wikipedias of 13 different languages with more than 4,000 active contributors of articles on ~9000 biological species that are present in at least 20 languages. We found that over time articles come to share more headers, particularly so within their respective categories. This can be seen as a pattern of standardization – however, this pattern can be detected to a different degree in the different Wikipedias (e.g. German around 2x more than English).

We also estimate the explanatory value of different social learning mechanisms in the process, through which the choice of headers in a new article may be dependent on the prevalence of headers in the existing population. We contrast the mechanisms of simple copying, popularity bias, prestige bias, recency bias and a time-invariant content bias. While they all provide a fairly good match due to equifinality and correlations between measures, simple copying seems to provide the best approximation in this case.