# EVOLANG DATABASE: A COMPREHENSIVE SOURCE OF INFORMATION ON THE SCIENCE OF LANGUAGE EVOLUTION COMMUNITY

ALEKSANDRA PONIEWIERSKA[*1], ALEKSANDRA SZCZEPAŃSKA[1], YEN YING NG[1], MARTA SIBIERSKA[1], MAREK PLACIŃSKI[1], PRZEMYSŁAW ŻYWICZYŃSKI[1] and SŁAWOMIR WACEWICZ[1]

[*]Corresponding Author: maestro.aleksandra@gmail.com

[1] Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

## 1. Introduction

In recent years, language evolution has further consolidated its identity as a distinct field of research (see e.g. Dediu & de Boer 2016; Kirby 2017; cf. the "Science of Language Evolution", Żywiczyński 2018) and has become the object of interest of metascientific analyses (see e.g. Bergmann and Dale 2016; Roberts et al. 2020). To contribute to this trend, we created a comprehensive database of information on the Evolang conference series (available at http://sle.cles.umk.pl/), as representative of the language evolution academic community. In what follows, we discuss the applicability of this database for scientometric research and present a case study of the academic profiles of those researchers involved in Evolang who self-identify with interests and disciplines most commonly associated with language evolution.

## 2. Data collection

The database contains information on Evolang abstracts (n = 1085) and Evolang contributors (n = 1401) over the period of 9 most recent iterations of the conference (2004-2020). The database was created in four steps: (1) A list of contributors was compiled based on automatic text mining of the Evolang books of abstracts. (2) Basic bibliographic information was collected from the proceedings volumes through automatic data collection, which allowed us to

identify the author profiles in Google Scholar (GS). Despite certain drawbacks, most importantly incomplete coverage of author profiles, GS has already been acknowledged as a useful tool in scientometric analyses (e.g. Gusenbauer 2019), and it proves particularly informative in studying self-reported data, such as one's discipline or research interests (see below). (3) Each author's profile was supplemented with data from GS (automatic data collection). The data collected includes self-reported affiliation and research interests, the year of the author's first publication shown in GS, as well as scientometric information such as h-index, i10-index, and citation count. (4) All the data collected was verified by three independent experts (manual data verification).

### 3. Database design

The database is organized into four tables: (1) "authors", including all Evolang contributors, (2) "publications", including bibliometric information on all Evolang abstracts, (3) "references", including full references used in the abstracts, and (4) "texts", including all the abstracts themselves. Tables (1), (2), and (3) are available in the user interface view of the database (see: http://sle.cles.umk.pl/). The data is fully interoperable and reusable, and in the future extended implementations it will be supplemented with additional information, such as on the geographical distribution of the authors.

| *authors* | | *publications* | |
|---|---|---|---|
| *field* | *source* | *field* | *source* |
| affiliation | GS | author(s): first and last names, affiliations | B |
| country | GS | excerpt | B |
| ORCID | M | references: all references used in the abstract | B |
| discipline | GS | citations: all citations used in the abstract | B |
| first Evolang contribution | B | type of document: short/long abstract | B |

| | | | |
|---|---|---|---|
| number of Evolang contributions | B | type of presentation: regular talk/plenary | B |
| first scientific publication | GS | bibliographic information | B |
| h-index (5-year and overall) | GS | | |
| i10 index (5-year and overall) | GS | | |
| citations | GS | | |

Table 1. A summary of the fields available in tables (1) and (2) of the database, and the relevant sources of data, where M stands for manual, GS stands for Google Scholar, and B for books of abstracts. Table (3) corresponds to the "publications/references" field; table (4) is partly available through "excerpts".

## 4. Analyses

To exemplify the type of questions that can be addressed with the current implementation of the database, we present an analysis of the self-identification of the Evolang contributors with a number of labels most often used to denote language evolution as a field of research. 642 authors in our database had identifiable GS profiles, with a total of 916 tokens of associated discipline/field labels (self-reported in GS as "Areas of interest"). Here, we looked into five labels: "language evolution", "evolutionary linguistics", "evolution of language", "language origin(s)" / "origin of language", and "biolinguistics" (see Table 2 below). This analysis helps us explore the subtle connotational differences between those five labels that are taken to be near-synonymous or overlapping. Of those five, "language evolution" is by far the most frequently used label by Evolang contributors, whereas "evolutionary linguistics" appears to be preferentially used by more senior and more accomplished researchers, oriented towards more computational research, and more strongly involved in the Evolang conference (of the total of four existing Google Scholar profiles with this label, three belonged to Evolang contributors). Perhaps surprisingly, only three Evolang contributors report "biolinguistics" as their area of research in their GS profile.

| | *language evolution* | *evolutionary linguistics* | *evolution of language* | *language origin(s), origin of language* | *biolinguistics* |
|---|---|---|---|---|---|
| 1. Number of authors | 69 | 15 | 21 | 4 | 3 |
| 2. Proportion Evolang to GS | .570 | .750 | .636 | .333 | 0.125 |
| 3. Other most common labels | cultural evolution, cognitive science, linguistics | computational linguistics, cognitive science | cognitive linguistics, comparative cognition | language evolution, linguistics | NA* |
| 4. Per cent long papers | 17.85 | 31.67 | 23.15 | 30 | 33.33 |
| 5. Seniority (years since first GS publication) | 15.94 | 26.27 | 18.63 | 16 | 16.33 |
| 6. Mean citation count | 1494.66 | 4825.4 | 2436.74 | 464 | 899 |

Table 2. A summary of the analysis of selected Google Scholar labels (self-declared "areas of interest") of Evolang contributors. Row 1 states the total number of Evolang authors for a given GS label. Row 2 states the share of Evolang authors among all GS profiles for that label. Row 3 states the other GS labels most commonly declared by Evolang authors with a given label. Row 4 states the proportion of papers (6-8 pages) to abstracts (2 pages) submitted by Evolang authors; we assumed the submission type to reflect the distinction between review / theoretical papers (typically longer) and reports of empirical results (typically shorter). Row 5 reports author seniority, operationalised as the time of active publishing, counting from the first publication listed on GS. Row 6 reports GS citation count.

## 5. Conclusions

As the field of language evolution research has rapidly grown and now constitutes to develop as a broad and highly interdisciplinary field, researchers have expressed the need to seek new methods to systematize and explore the scientific production in this field (Bergmann & Dale, 2016; Roberts et al. 2020). We believe that the database reported in this contribution provides a valuable

tool and resource to this end, with a range of theoretical and practical applications. In future work, we plan enriching the database with information from complementary resources, which would further extend the range of possible analyses. One example is the inclusion of GS citation data on individual publications, which will make it possible to identify the classic references (i.e. the publications most often – and most interdisciplinarily widely – cited in the Evolang proceedings), and address questions on the nature of the contribution of the different disciplines, research centers and author networks and its dynamics across the successive iterations of the Evolang conference. Another interesting direction is integrating our database with thematically related databases – in particular, the CHIELD database (Roberts et al. 2020), which contains entries for a large number of Evolang articles, with manually coded information such as on the methods used in the reported studies (e.g. “experiment”, “observation”) and the stages of language emergence to which they refer (e.g. “biological evolution”, “language change”).

**Acknowledgements**

This research was supported by the Polish National Science Centre under grant agreement UMO-2019/34/E/HS2/00248.

**References**

Bergmann, T., and Dale, R. (2016). A Scientometric Analysis Of Evolang: Intersections and Authorships. In S. G. Roberts et al. (eds) *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. New Orleans: Evolang Scientific Committee. Availa.

Dediu, D., & Boer, B.G. (2016). Language evolution needs its own journal. *Journal of Language Evolution, 1,* 1-6.

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118(1),* 177-214.

Kirby, S. 2017. Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review 24(1)*, 118-137.

Roberts, S. et al. (2020). CHIELD: The Causal Hypotheses in Evolutionary Linguistics Database, *Journal of Language Evolution, 5(2)*, 101-120, DOI: 10.1093/jole/lzaa001.

Żywiczyński, P. (2018). *Language Origins: From mythology to science*. Berlin: Peter Lang.