Journal of Language Evolution, 2024, **XX**, 1–17
https://doi.org/10.1093/jole/lzae003
Advance access publication 2 July 2024
Research Article

OXFORD

The evolution of evolutionary linguistics

Poonam Brar[1], **Chico Q. Camargo**[1,2,3,4,*]

[1]Department of Computer Science, University of Exeter, EX4 4RN, Exeter, United Kingdom
[2]Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, United Kingdom
[3]Department of English Language and Literature, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea
[4]Alan Turing Institute, 96 Euston Rd., London NW1 2DB, United Kingdom
[*]Corresponding author. Department of Computer Science, University of Exeter, EX4 4RN, Exeter, United Kingdom. E-mail: f.camargo@exeter.ac.uk

Associate Editor: Dr. Seán Roberts

This paper presents a scientometric study of the evolution of evolutionary linguistics, a multidisciplinary field that investigates the origin and evolution of language. We apply network science methods to analyse changes in the connections among core concepts discussed in the Causal Hypotheses in Evolutionary Linguistics Database, a searchable database of causal hypotheses in evolutionary linguistics. Our analysis includes a multipartite network of 416 papers, 742 authors, and 1,786 variables such as 'population birth rate' and 'linguistic complexity'. Our findings indicate a significant increase in the size of concept networks from 1886 to 2022, providing an account of the growth and diversification of evolutionary linguistics as a field. We describe eight major clusters of concepts, and characterize the connections within and between clusters. Finally, we identify hypotheses cutting across clusters of concepts that have a high-betweenness centrality, implying that they might have a higher impact on the field if proven right (or wrong). Furthermore, we discuss the role of databases in cultural evolution and scientometrics, emphasizing the value of interdisciplinary connections and the potential for further cross-disciplinary collaboration in the field of Evolutionary Linguistics.

Keywords: evolutionary linguistics; network analysis; community detection; collaboration network; crowdsourced data.

Introduction

One of the boons of the 21st century was the rise of Network Science as a discipline made possible through the digitization of copious amounts of print data (Barabási 2016). Even though networked structures are arguably as old as life, advanced technology and digital data have made it possible to map and analyse networks in almost all fields of knowledge, be it business, nature, or social phenomena (Barabási 2016). Today, in science, we have digitized databases for authors, their research papers, citations, patents, and grants in numerous journals and online resources (Fortunato et al. 2018). The availability of massive data makes it possible, for instance, to map citation networks to understand the connection of scientific topics, or to use author networks to visualize the contribution of different authors to a scientific field. A quantitative understanding of such networks can yield important insights into subsequent scientific discoveries and even help design policies to advance science (Fortunato et al. 2018). It would also help understand similar scientific breakthroughs across different disciplines (Guevara et al. 2016). For instance, a researcher from a specific academic field can cite papers from other academic areas in her reference section. The paper cited from a different discipline would connect to the documents cited from her field, ultimately creating an interdisciplinary research network (Guevara et al. 2016). Identifying the patterns in these graphs using statistical and computational methods can give us valuable insights into the behaviour of entities under study (Newman 2010).

This paper focuses on implementing the methods and concepts of network science in a specific field of science: Evolutionary Linguistics, which in its simplest form, is the study of the evolution of language (MacMahon and MacMahon 2012). A comprehensive analysis of how different hypotheses in Evolutionary Linguistics are connected, looking at when theories are in agreement, contradiction, or simply when theories do not interact with each other, should provide insights

into the kind of research that might be more relevant for future works.

In this work, we perform network analysis methods on the data available in the open-source database Causal Hypotheses in Evolutionary Linguistics (CHIELD). CHIELD, created by Roberts et al. (2020), represents different hypotheses about language change and evolution on a single network where 'variables' represent the essential ideas around which the research revolves, connected through causal links. Fig. 1 shows an example of causal relations among variables discussed by Frank and Smith (2020) in their research. Frank and Smith (2020) argue that the birth rate in a human society will determine its population size and the proportion of young learners in that community, which in turn should determine linguistic complexity. In their paper, the authors argue that more young learners in a population are likely to result in lower linguistic complexity (Frank and Smith 2020).

The online platform accompanying CHIELD allows for informative graph visualizations, connecting variables from different papers according to different criteria. However, if one were to represent all 1,700 variables and 3,400 links together on a single graph, the task of analysing the full network and identifying how evolutionary linguistics theories relate to each other becomes intractable without the use of network science tools. By applying these tools to analyse this large network of variables, papers, and authors, we aim to provide deeper insight into the evolution of evolutionary linguistics.

In this paper, we offer a comprehensive overview of the origins and evolution of evolutionary linguistics, delving into a century's worth of research. We analyse the variable–variable network to chart the field's development and key themes, which we operationalize as communities within the network, as well as the co-authorship network emerging from CHIELD, which we use to shed light on significant contributions from authors to the growth and advancement of this field. Drawing on CHIELD data, we identify crucial variables, authors, and papers across pivotal periods in evolutionary linguistics. Lastly, we explore potential future directions for evolutionary linguistics, by identifying hypotheses of high edge betweenness centrality which also cut across communities of variables, which we identify as potential cases of future work which could lead to a high impact in the field.

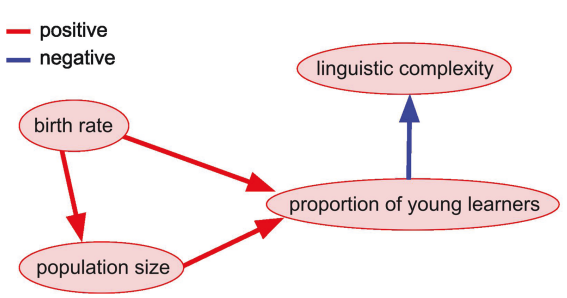

Figure 1 Snapshot of variables in CHIELD. Nodes represent variables, and (red/blue) edges represent (positive/negative) impact of one variable on another. Color version of the figure is available online.

Background and related work

Evolutionary linguistics

When and how did language originate? Why is *Homo sapiens* the only species that possesses a complex linguistic structure? How does language get transmitted to new learners and eventually modified in the process? These are the kinds of questions that evolutionary linguistics aims to answer (Ke and Holland 2006; McMahon and McMahon 2012; Gong et al. 2013; Nölle et al. 2020) and provide us with hypotheses, experiments, and theories about the evolution of language (Kirby and Christiansen 2003; Számadó and Száthmary 2006). According to Kirby and Christiansen (2003), language is the hardest problem in science because of multiple theories about language origin and evolution—some in consensus, some controversial, and some completely different perspectives.

There has been no shortage of discussion about how to define and delimit the field. Haspelmath (2020) points out how the term 'language' has been used to 'refer to particular languages (sets of conventions used by particular speech communities), or to the use of a language in speech, or to the entire domain of phenomena related to language use and language systems' (Haspelmath 2020). The author then proceeds to point out how the expression 'language evolution' might refer to the evolution of the capacity for languages, or to the evolution of a language system over time—what other authors might call diachronic change. This is perhaps unsurprising, considering how many fields are touched by language, which is the basis of so much of human interaction. Today, the field of language evolution/the evolution of language/evolutionary linguistics is a multidisciplinary research field that makes use of hypotheses, experiments, and simulations across linguistics, archaeology, biology, neuroscience, behaviour, mathematics, and computer science (Christiansen 2003; Gong et al. 2013; Nölle et al 2020).

Scott-Phillips and Kirby (2010) have devised four stages of language evolution: pre-adaptation, co-evolution, cultural evolution, and language change; Roberts et al. (2020) have also used this categorization to classify variables in these stages.

The initial stage, often referred to as pre-adaptation, involves biological and evolutionary changes that,

while not originally related to language, laid the groundwork for complex communication. For example, bipedalism, a significant evolutionary milestone, inadvertently influenced the structure of the vocal tract, facilitating more varied vocalizations (Hurford 2003; MacWhinney 2008). Early humans also developed sophisticated cognitive abilities, allowing for nuanced social interactions and the emergence of complex communication systems (Hurford 2003).

The transition from individual capabilities to shared communication systems marks the co-evolution stage. This phase suggests a symbiotic development between the inherent biological predispositions and the cultural practices surrounding language use. The ability to remember and replicate tool-making processes, for instance, paralleled the retention of communicative cues, paving the way for language to take root within early human communities (Hurford 2003).

Cultural evolution, the subsequent stage, investigates how language is acquired and transmitted across generations. Children learn language not merely by mimicking adults but through a dynamic process of engagement with their linguistic environment, which is shaped by various factors including cultural norms and social structures (Tomasello 2003; Hoff 2006). This stage underscores the complex interplay between innate linguistic capacities and the sociocultural context in which language learning occurs.

Language change, the final stage discussed, highlights the evolving nature of language structures. Languages are not static; they undergo transformations influenced by a multitude of factors, from phonetic shifts to changes in grammar and syntax. These alterations can lead to language diversification or, in some cases, language extinction. An example of this could be since an increase in the youth population causes language simplification (Frank and Smith 2020), we observe that the verb 'going to' gets modified to 'gonna' (Croft 2000) in the informal manner of speaking, adding another word that conveys the same meaning (Croft 2000).

Understanding the complexities of language evolution requires a multidisciplinary approach, drawing on evidence from fields as diverse as archaeology, genetics, neuroscience, and psychology. Each discipline contributes unique insights, from the study of ancient artefacts and fossils that offer clues to early human communication, to modern experiments that explore the cognitive underpinnings of language acquisition and use (Gardner 1983; Davidson 2003; Hauser et al. 2003; Lieberman 2003; Cheney and Seyfarth 2005; Tattersall 2014).

The causal hypotheses in Evolutionary Linguistics database

The field of Evolutionary Linguistics comprises multiple theories, some clashing with each other, some complementing each other, and some completely disjointed. Roberts et al. (2020) highlight these problems in their study and propose a solution: a database with all hypotheses relating to the evolution of language, drawing from multiple disciplines and placing all hypotheses in one searchable database. CHIELD is a database containing information on over 400 research papers published since 1886, containing not only metadata describing each paper, such as author list and year of publication but also the essential concepts discussed within them, which the database refers to as 'variables'. Within CHIELD, variables are connected based on cause and effect relationships, such as 'a change in X causes a change in Y' or ''X exerts an evolutionary selection pressure on Y'. Roberts et al. have defined two criteria for link categorization: stage (evolutionary linguistics stages by Scott-Phillips and Kirby 2010) and type of study (experiment, review, model, simulation, statistical, qualitative, logical, hypothesis, or something else). The database also describes a note for each link to provide more context about the variables' relationships. The data within CHIELD is contributed by researchers themselves or by contributors engaged in related fields.

In a recent study of relevance for this paper, Wacewicz et al. (2023) undertook a scientometric analysis of the evolution of language following inspiration from Bergmann and Dale (2016). However, their approach differs as they constructed a database spanning sixteen years from the EvoLang conference, instead of using CHIELD. Their study employed natural language processing techniques to identify the frequency of common topics among the papers, along with spatial analysis and other approaches. In contrast, our focus centres on an established database where topics are meticulously curated by field experts, forming the basis for our analysis.

All results below use data collected from CHIELD in March 2022, containing 416 research papers, 742 authors, and the pairwise causal relationships among the 1,786 variables in the field of Evolutionary Linguistics as annotated in CHIELD. All code for analysis was written in Python, and statistical analysis was conducted using Gephi 0.92 (Bastian 2009), open-source software for network exploration, manipulation, and statistics.

Network science

Within-network science, the typical object of study is a network with nodes that represent the entities under study, and edges depict the pairwise associations between them (Barabási 2016; Newman 2010). The relationship between any two nodes is either directed, where each edge is traversed in only one direction (e.g. when one author cites another), or undirected, where

the connection between the nodes is two-way (e.g. when representing two people sharing authorship on a paper) (Newman 2010; Barabási 2016).

In network science, a connected component is a subnetwork in which a path connects every pair of vertices, and which is not part of any larger subnetwork with the same property. A network can have multiple connected components, i.e. multiple groups of connected nodes in a network that are not connected to any other nodes in the network. The largest connected component of a network is called the giant component (Barabási 2016).

Within a network, we can identify multiple communities of nodes. We use statistical community detection methods to uncover the hidden structure of a network and to understand how the nodes within a network are organized and interconnected. These communities are characterized by a dense cluster of connections between the nodes within the group and a sparser set of connections between the nodes in the group and the rest of the network (Newman 2006). We have used the terms 'cluster' and 'community' interchangeably throughout the paper, implying the same thing. Various algorithms and techniques exist for this task, and the choice depends on the network's specific characteristics.

Methodology and results

Here, we build two networks from the CHIELD data: an undirected author–author network, where nodes represent authors and edges between them represent co-authored publications, and a directed variable–variable network, where nodes represent variables such as 'population size' and '[language] complexity', and edges represent the relations between such variables, including causal relations such as 'a change in population size causes a change in morphological complexity', as well as non-causal relations such as 'X and Y co-evolve'.

Figs. 2 and 5 respectively represent the largest connected components of the author-author network in CHIELD, containing 172 authors out of 742 in the database, and the evolution of the variable–variable network, described below. Different colours on the images indicate the communities obtained using the Louvain community detection algorithm, respectively representing groups of authors that collaborate with each other more often than with those outside that community and groups of variables that tend to be linked more often to each other.

The Louvain community algorithm consists of two phases: in the first phase, nodes are assigned to communities and evaluated based on modularity gain, while in the second phase, a new network is constructed, and the process is repeated to form giant nodes representing communities (Blondel et al. 2008). In rudimentary form, modularity is the difference between the number of edges that are a part of a community and the expected number of edges that might have been had the network formed randomly (Newman 2006).

In addition to Louvain community detection, we also tested the Label Propagation algorithm and the Girvan–Newman method, looking at multiple partition levels for the latter. All of them resulted in partitions with lower modularity than the one obtained using the Louvain method, which suggests the communities identified by the Louvain algorithm might be a more useful description of the networks in this study.

We also conducted a qualitative comparison of the communities identified by the Louvain algorithm versus those identified by the Girvan-Newman when partitioned into sixteen communities, which is the number of partitions that led to the largest jump in modularity values (see Supplementary Appendix). We find that the communities identified by the two methods are robust, with small variations between those found by one method and those found by the other. The Louvain method tends to produce broader, more encompassing communities, while the Girvan-Newman method seems to produce more specialized clusters. The differences might also reflect the methodologies themselves: the Girvan–Newman method, based on edge betweenness, might lead to more granular communities by iteratively removing high-betweenness edges, while the Louvain method, optimizing modularity, might result in larger, more inclusive communities by aggregating nodes that contribute to denser connections within communities rather than between them. The complete comparison is presented in the Supplementary Appendix.

One important point about community detection via modularity maximization is to not interpret it as the unambiguous truth about a network: modularity maximization has a characteristic scale, and tends to find communities of similar size, in particular with the same sum of degrees. This approach would perhaps not be suitable for a study focused on producing representative statistics of the whole field of evolutionary linguistics, or a study where community sizes were essential to the findings. In such a context, Bayesian algorithms for stochastic model inference would be more suitable (Peixoto 2023). However, for this study, modularity maximization is a useful feature, as it allows us to partition the network into roughly equal-sized parts.

We also measure how some authors act as bridges between communities using the authors' *betweenness centrality*, which quantifies the number of times a node acts as a bridge along the shortest path between two other nodes (Barabási 2016). This measure stands in contrast to centrality measures which emphasize different aspects of a node: such as degree centrality,

which literally measures the number of connections of a node, or closeness centrality, which measures how far a node is from other nodes, or eigenvector centrality, which is a measure of the importance of a node based on the importance of those they are connected to, and is usually seen as a measure of influence (Newman 2010). As nodes that are central according to one metric tend to be ranked high in centrality according to other metrics, we also provide the degree, closeness, and eigenvector centrality measures in the Supplementary Appendix, but we focus on the betweenness centrality in the main text, as this metric represents the role of a node as a connection between one part of the network and another. As Wacewicz et al. (2023) mention in their analysis of the EvoLang conference, betweenness centrality highlights 'brokers', i.e. authors who bridge gaps between various sub-disciplines (Wacewicz et al. 2023). In this study, the same applies to variables in CHIELD: variables with a high-betweenness centrality can be understood to be the bridges between multiple subfields.

Larger nodes in Fig. 2 have higher betweenness centrality. The most significant node according to betweenness centrality is Stephen C. Levinson, who lies on the maximum number shortest paths between any two nodes in the network, making it a very central author to evolutionary linguistics. This measurement also highlights this author's role in the author collaboration landscape on CHIELD, since he has collaborated with thirty-seven authors on nine papers.

When comparing our co-authorship network analysis to the work of Wacewicz et al., there are notable differences in the top influential authors identified. In their study, applying betweenness centrality measures on their dataset revealed Simon Kirby, Susan Goldin–Meadow, and Kenny Smith as the top three influencers (Wacewicz et al. 2023). In contrast, the same analysis applied to CHIELD identified Stephen L. Levinson, Dan Dediu, and Gerhard Jäger as the authors with the highest betweenness centrality. Despite their unique role in the CHIELD co-authorship network, we find that their centrality scores are within what would be expected by chance for this network: we confirm that by taking a bootstrap sample of 1,000 random networks with the same size and degree distribution as the CHIELD co-authorship network, measuring the betweenness centrality of all nodes, producing a distribution of the top ten betweenness centrality values for each network, and comparing them with the top ten authors from CHIELD. The top ten betweenness centrality values for CHIELD fall very much within the distribution produced from the bootstrap sample, with P-values ranging from $P = 0.008$ (rare, but still found within the bootstrap sample) to $P = 0.80$ (very common in the bootstrap sample). This is shown in the Supplementary Appendix, and discussed in more detail in the Discussion and Conclusion.

This difference in results between those obtained using the data collected from EvoLang by Wacewicz et al. and the one in CHIELD is expected: the former only looks at works presented at EvoLang, but is comprehensive within that sample, whereas CHIELD includes annotations regarding papers published in other venues, but does not have the same comprehensive coverage.

The evolution of CHIELD

The CHIELD database provides a window into the evolution of evolutionary linguistics. Fig. 3a shows the number of papers published each year in the database. In the early years (1886–1960) present in the database, publications are rare and far apart. From 1967, the frequency of papers over time increased, with one publication appearing almost every year in the database, and eventually multiple papers per year. This increasing trend hits its maximum at 66 published in 2018, followed by a decline after that year. The average number of authors per paper, shown in Fig. 3b, also grows over time, indicating how the field of Evolutionary Linguistics as captured by CHIELD has become more collaborative over time.

Fig. 4a shows the number of nodes in the largest component of the variable–variable network over time. It shows an expressive growth since 1995, indicating a more thorough coverage of all the variables discussed in the field of evolutionary linguistics. Fig. 4b shows the evolution of the average degree and density of the network. Degree is defined as the average number of connections per node in the network, and density is the ratio between the number of connections in the network and the total number of possible edges for a given network. Together, both plots show how the network grows over time, in that more variables are connected to each other, but also that most variables are connected to only a few.

Table 1 shows the evolution of a few key network statistics for the variable–variable network, for three time periods: 1886–1984, when a maximum of one paper was published each year, 1985–2005 when more than one but less than ten papers were published, and 2006–2022 when more than ten papers were published each year. It is important to notice that this variation in the number of papers in the database is partly due to annotation biases in CHIELD itself: since our findings represent the data on the CHIELD website, the trends in Fig. 4 and Table 1 only describe the papers and variables identified by the CHIELD contributors.

Fig. 5a–c shows three snapshots of the network at the three intervals described above. During 1886–1984, in Fig. 5a, we observe variables from eleven different

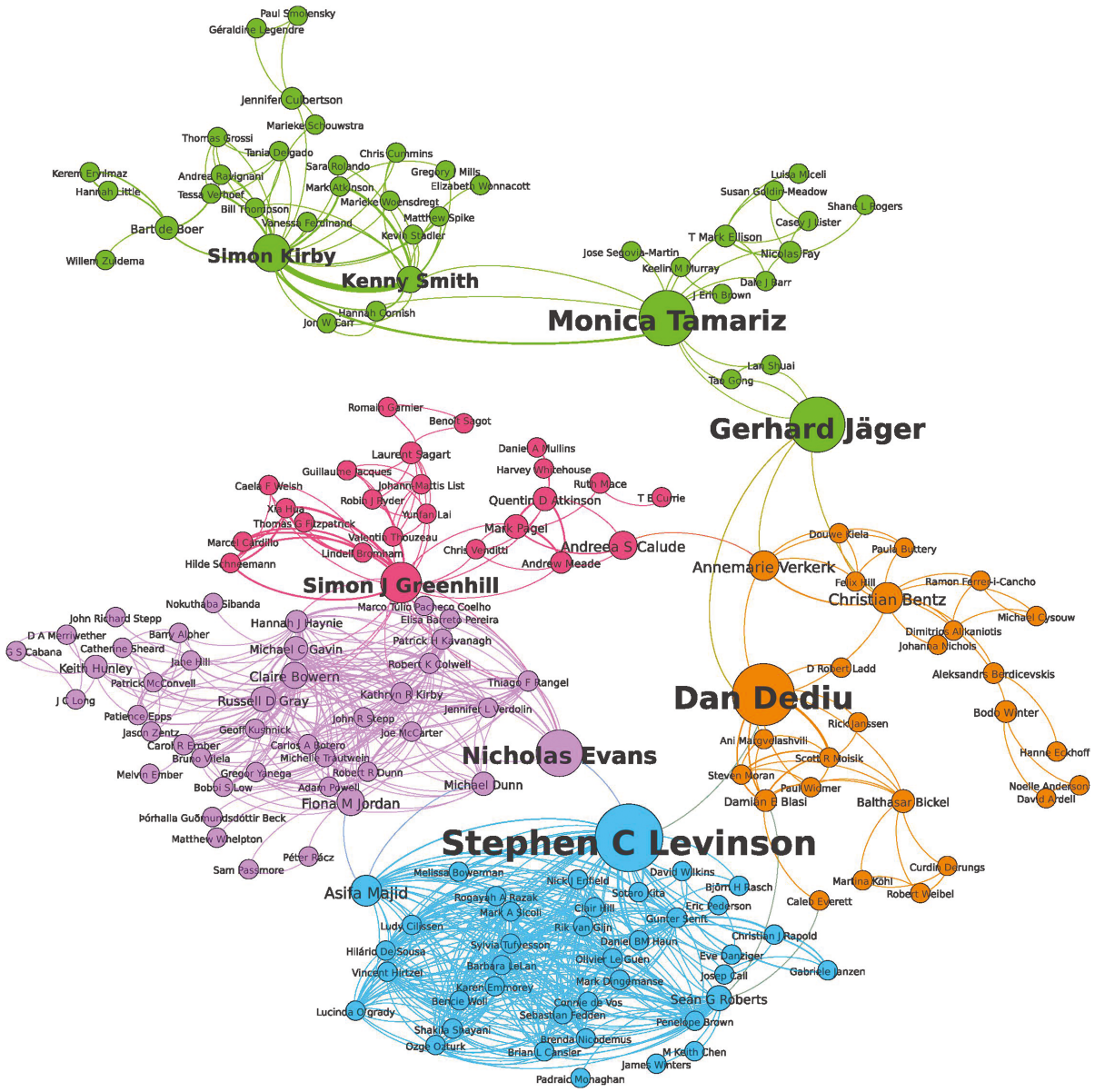

Figure 2 The giant component of the co-authorship network in CHIELD. In the figure, every colour represents a community of authors. Larger nodes and node labels indicate nodes of higher betweenness centrality.

communities, but they are mostly disconnected. The connected variables mainly belong to the same paper. Variables such as *language*, *population size* and *latitude*are introduced during this period. From 1886 to 2005, in Fig. 5b, we observe that denser connections build up during this interval. The variables connect well with each other, and more communities take shape. Fig. 5c shows the final full-scale network that represents the present-day connections of variables in the CHIELD database.

The giant component for the variable–variable network in the complete time range, i.e. 1886–2022, contains 1,660 nodes and 3,247 edges. The average degree for this network is 1.99, indicating that the average variable is connected to two more variables on the network. It is, however, an average: while many variables show only a single link to another variable in the dataset, the variable 'language' holds 114 causal links to other variables, found in a total of 40 papers. This growth can also be seen in the network diameter, i.e.

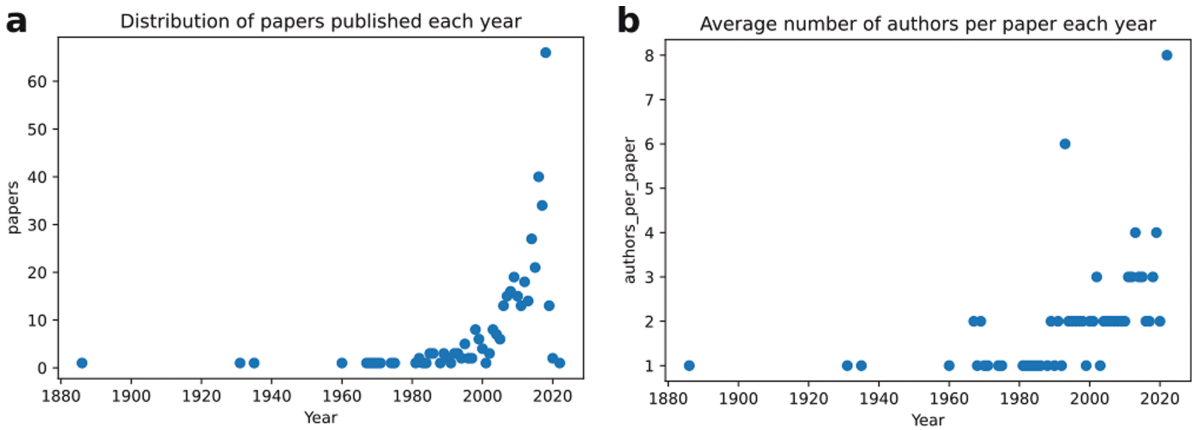

Figure 3 (a) Number of papers published each year as recorded in the CHIELD database, (b) average number of authors per paper.

the length of the longest shortest path between two nodes on the network, which goes from five in 1985 to twenty two in 2022.

To capture the importance of different nodes in the variable–variable network, we computed the betweenness centrality of every node. We observe the following top ten variables in our network, as also labelled in Fig. 5c, listed below, from most central (highest betweenness centrality) to least central (lowest betweenness centrality):

1. Population size
2. Language
3. Cooperation
4. Group size
5. Language diversity
6. Brain size
7. Vocal learning
8. Self-domestication
9. Compositionality
10. Protolanguage

Fig. 5c shows that some variables have formed fan-like structures towards the periphery of the network. Such structures occur when multiple variables connect to a single variable used in the same or different papers. The dashed red circle on the upper side of the network highlights one distinct structure in the community *Neurolinguistics and Language Disorders* (communities are explained in detail below). In that fan structure, one of the main variables—*brain size*—lies in the inner part of the network, connected to *cognitive deficits*, *speech/ language problems*, and *self-domestication* in the dark blue community next to it. These three variables then connect with a bundle of nodes belonging to genetics and neuroscience concepts, such as *neural crest cells*, as well as twenty one variable nodes labelled with different genes, and other neuroscience variables. We can thus conclude that *brain size* is the central node that connects these genetics/neuroscience nodes with other disciplines.

At the bottom of Fig. 5c, another dashed red circle highlights a thick edge representing a strong link between the variables *obligatory grammatical distinction: future time reference* and *future discounting*. The thick edge indicates that the connection traverses multiple times between these nodes; this is the case because Chen (2013) has provided multiple criteria for the relationships between these two variables based on their investigation. In this particular case, *obligatory grammatical distinction: future time reference* refers to the languages with a strong future time reference (FTR), and *future discounting* refers to the economic outcome of this structure. The causal relationships between the two variables depend on the experiment result, for instance, whether families that speak FTR languages save more money, smoke more cigarettes, or are more prone to weight gain. Note that this is quite an exceptional case, with so many links coming from the same paper; most of the time, the multiple mentions of a causal link will come from different publications, and therefore different contexts where the link is inferred or theorized.

Communities in the variable–variable network

In Fig. 5c, the variables in the giant component of the network are clustered into twenty communities. In our analysis, we will focus on the top eight communities, which contain just under 60% of all variables. We have named those communities based on the highest-degree variables present in each community identified by the Louvain method. Since communities vary in size and scope, the names used here are mostly for description purposes: some communities of variables mostly

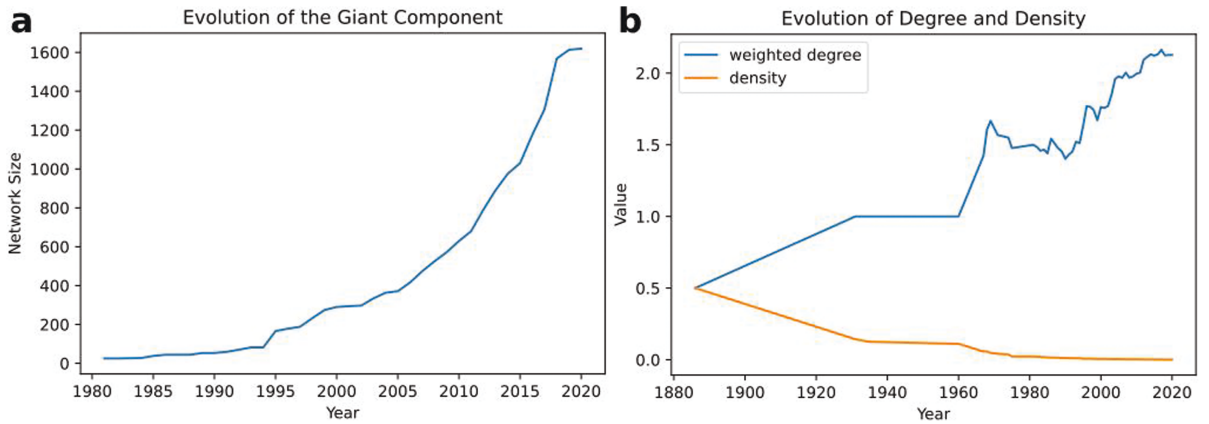

Figure 4 The change in the size of giant component over time.

Table 1. Key network statistics over time for the variable–variable network. The statistics are shown for the total (cumulative) network in each period, meaning that they include papers published until then.

	1886–1984	1886–2005	1886–2022
Number of nodes	25	281	1,660
Number of edges	34	397	3,247
Average degree	1.36	1.41	1.99
Network diameter	5	13	22
Average path Length	2.59	4.74	6.14
Graph density	0.057	0.005	0.001

describe a stage of the evolutionary history of language (e.g. Community 1: *Early Language Evolution*), while others describe broad areas of research (e.g. *Sociolinguistics and Language Complexity*). The list of all community names and key variables, along with the sixteen communities identified by the Girvan–Newman method for comparison, are listed in the Appendix. Community names are here presented in decreasing order of community size:

1. Early Language Evolution
2. Sociolinguistics and Language Complexity
3. Anthropological Linguistics and Kinship
4. Cognition and Social Interactions
5. Linguistic Diversity and Geography
6. Lexical Studies and Psycholinguistics
7. Neurolinguistics and Language Disorders
8. Cultural and Ritualistic Elements

The *Early Language Evolution* and *Sociolinguistics and Language Complexity* communities have the most variables—12.3% and 10.9% respectively. This section describes the relationship between those communities and other main communities in the network.

Fig. 6 shows the bridges between *Early Language Evolution* and *Sociolinguistics and Language Complexity*. In total, it shows twelve edges between the variables of the two communities. The four main variables from the two communities are—*language, cooperation, population size*, *and linguistic variation between groups*. These variables are the main connection points between both communities, forming bridges with other variables. Among the four main variables, only *cooperation* and *linguistic variation between groups* are bridged directly, while the other main variables have at least two edges connecting them. We also observe smaller nodes connected directly, such as *regularization bias* and *domain-specific systems*.

Fig. 7 shows an example of extensively connected communities. The primary variable, *cooperation*, connects to thirteen variables from the *Cognition and Social Interactions* community. Furthermore, the other primary variable, *language*, in the *Early Language Evolution* community, connects with eight variables in the *Cognition and Social Interactions* community. In the *Cognition and Social Interactions*community, the variables *social interaction, group size, brain size: neocortex* are the most connected with the *Early Language Evolution* community.

Fig. 8 shows that the connections between the *Cognition and Social Interactions* and *Sociolinguistics and Language Complexity* communities are less dense than community relationships in the two previous examples. We have twelve bridges between the two communities, and not all key variables in each community are directly connected. Only the main hubs from each community—*population size* from the *Sociolinguistics and Language Complexity* community and *group size* from the *Cognition and Social Interactions* community—are connected.

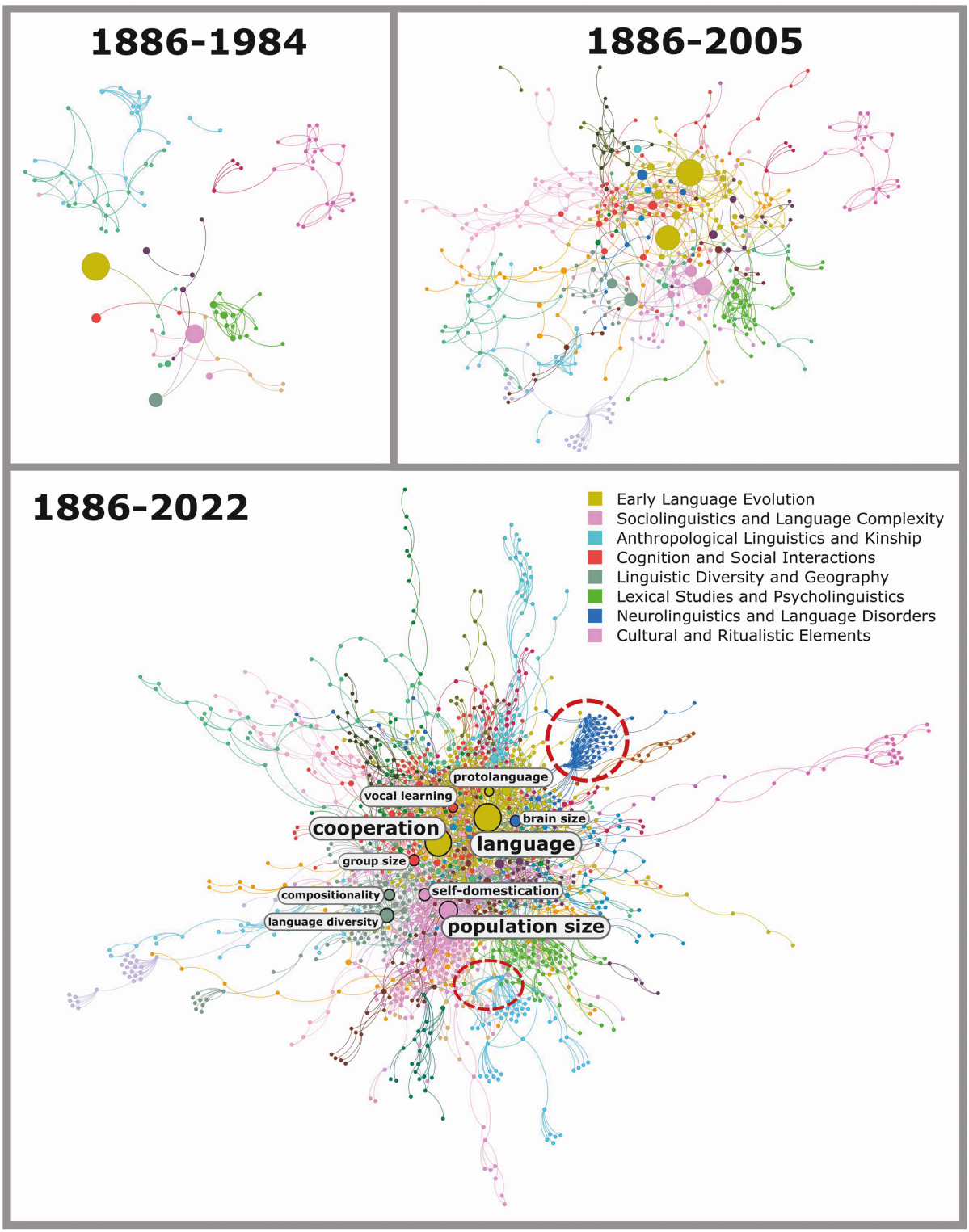

Figure 5 The variable–variable network from (top left) 1886 to 1984, (top right) 1886 to 2005, and (bottom) 1886 to 2022. Nodes represent individual variables, connections represent causal links between them and colours represent communities in the network, that is, groups of densely connected variables.

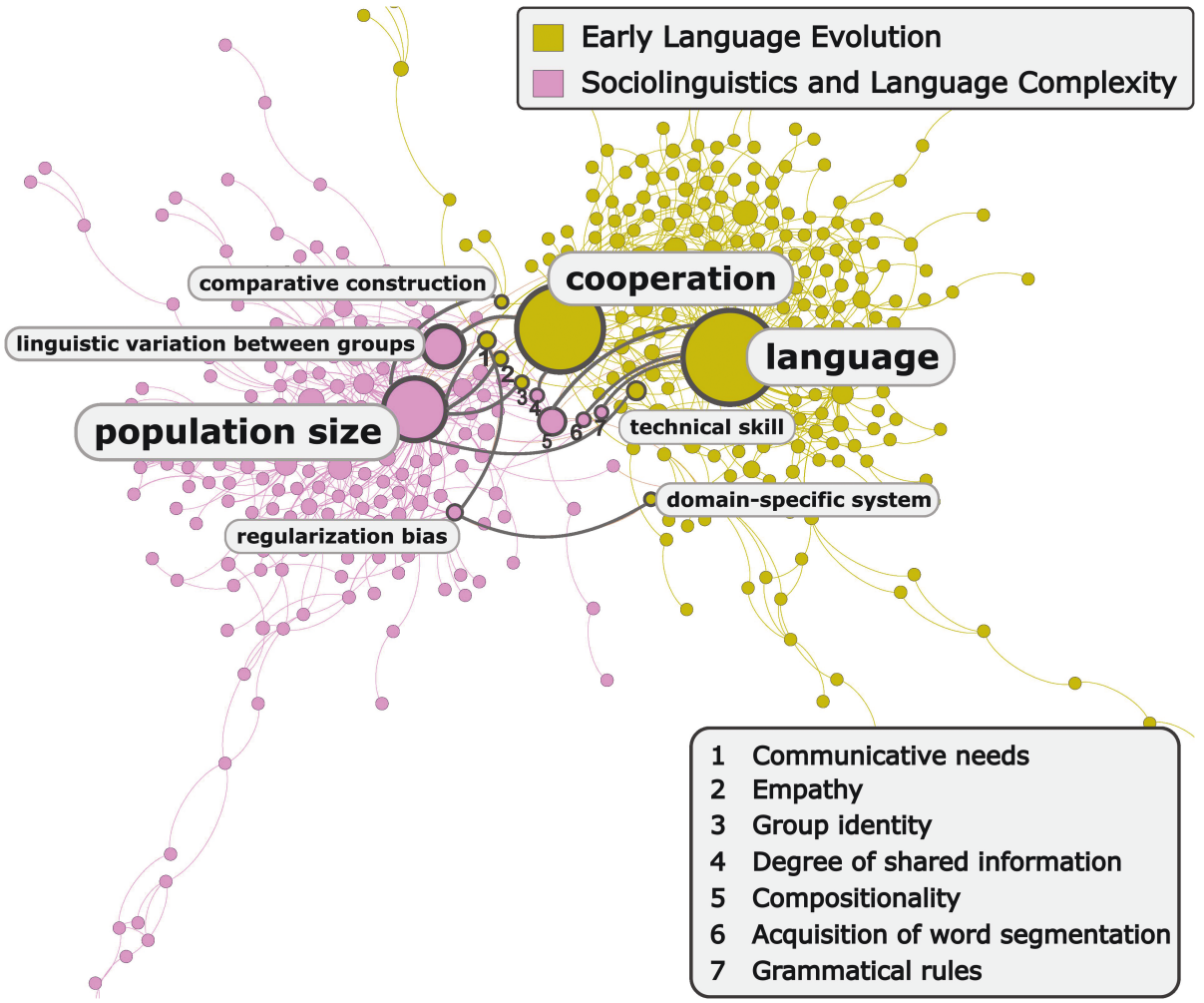

Figure 6 Community bridges graph for *Early Language Evolution* and *Sociolinguistics and Language Complexity* communities.

It is worth noting that the examples above are particular cases, as some communities are not connected at all. For instance, the *Neurolinguistics and Language Disorders* community is not connected directly with the *Cultural and Ritualistic Elements* or the *Lexical Studies and Psycholinguistics* communities. The heatmap in Fig. 9 shows a complete picture, indicating the number of causal links between every pair of communities in the variable–variable network. The strongest observed relationship is between *Early Language Evolution* and *Sociolinguistics and Language Complexity*, shown in Fig. 6. The *Early Language Evolution* community is also strongly connected to *Sociolinguistics and Language Complexity* and *Neurolinguistics and Language Disorders*, making it the community with the most connections to other communities.

Hypotheses bridging across communities

Even though community detection algorithms such as the Louvain method are well-suited to identify sets of nodes (variables, in this case) that share more links between them than the average pair of nodes in the network, they do not discern between *types* of connections. In this section, we look at the distribution of different types of edges in the variable–variable network, such as *Experiment*, *Hypothesis*, and *Model*.

The composition of the links connecting variables within the same communities is displayed in Fig. 10. In the figure, each colour represents one type of edge, indicating different types of evidence for a relationship between two variables, such as a literature review (*Review*) or the result of a computational simulation (*Simulation*), as well as logical connections and hypotheses (respectively, *Logical* and *Hypothesis*).

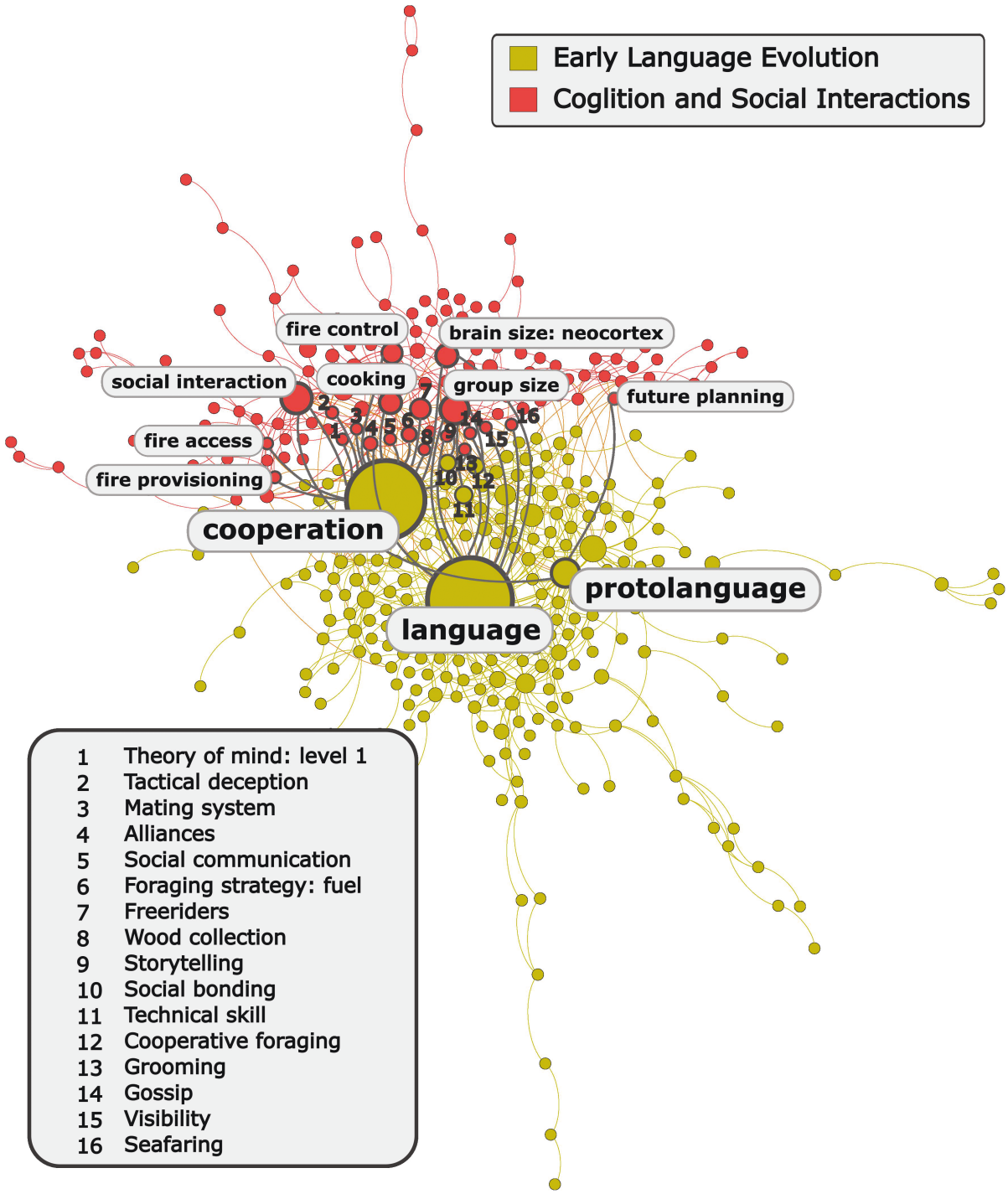

Figure 7 Community bridges graph for *Cognition and Social Interactions* and *Early Language Evolution* communities.

The *Early Language Evolution* and *Sociolinguistics and Language Complexity* communities are the ones with the highest number or within-community links, and also the highest number of within-community hypotheses, as indicated by the wide orange bars in both rows. At the same time, the two communities show very

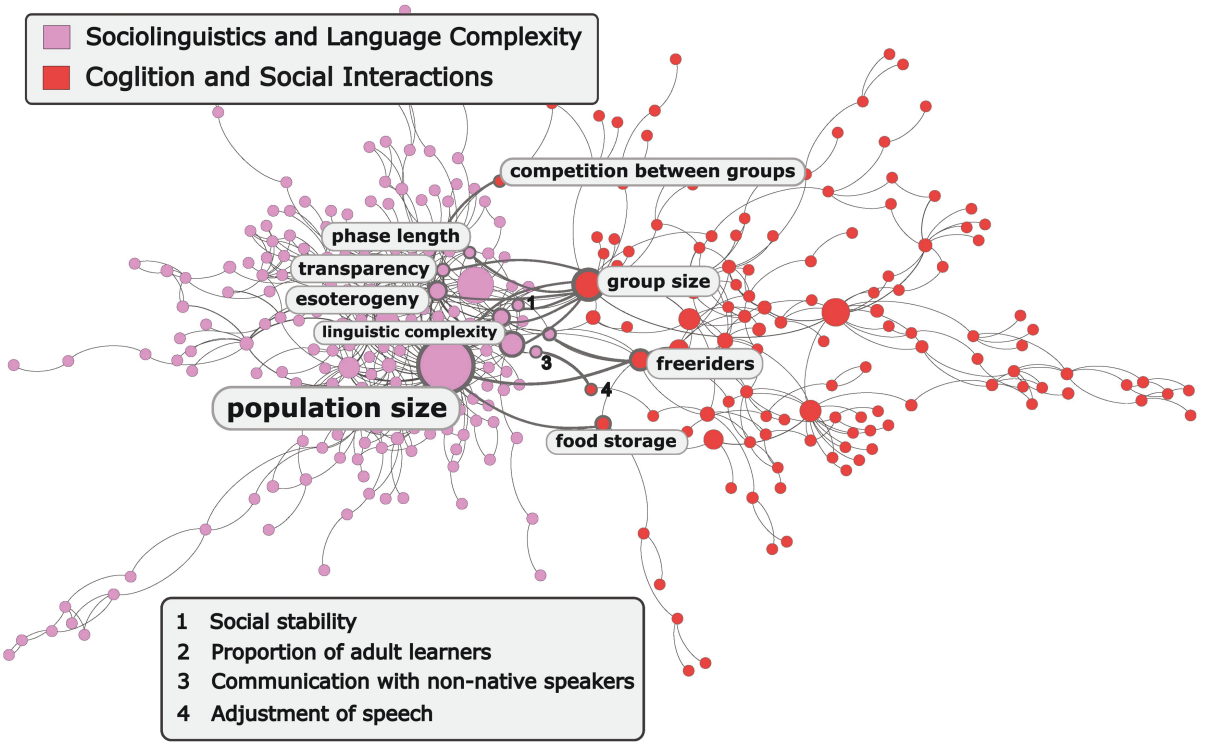

Figure 8 Community bridges graph for *Cognition and Social Interactions* and *Sociolinguistics and Language Complexity* communities.

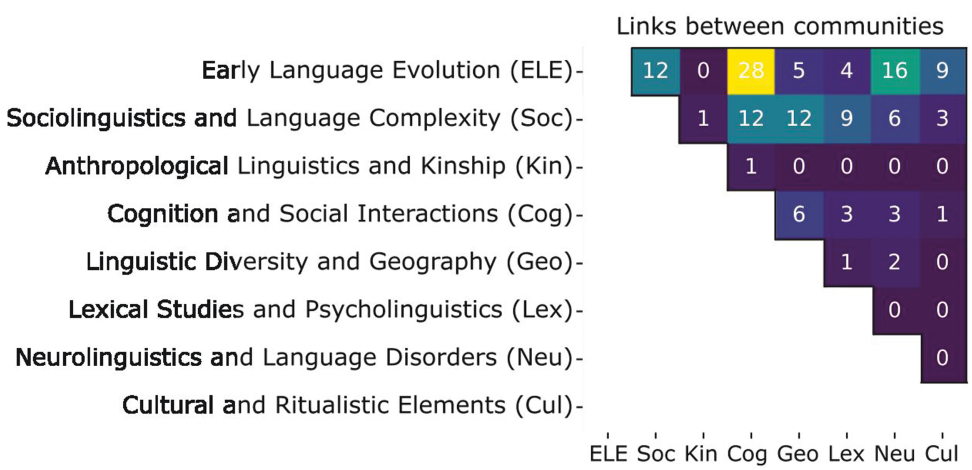

Figure 9 Heatmap showing the number of links between communities, that is, the number of variables shared between each pair of communities.

distinct distributions of link type: the *Sociolinguistics and Language Complexity* community presents a much more diverse profile, with *Review* edges being almost as frequent as *Hypothesis* edges, and a relatively high frequency of edges describing other forms of evidence, such as *Qualitative*, *Statistical*, and *Experiment*. In the *Early Language Evolution* community, by comparison, *Hypothesis*-type edges account for over 60% of the within-network links. This difference occurs across other communities in different degrees, and can be interpreted in multiple ways: it may suggest that communities with a diverse edge profile such as *Sociolinguistics and Language Complexity* are more well-established and empirically validated, whereas

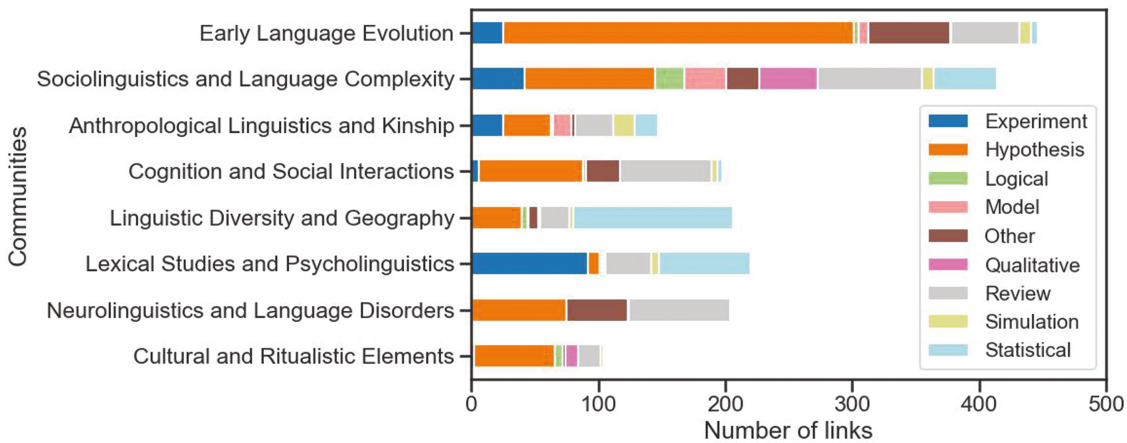

Figure 10 Composition of the links connecting variables within the same communities. In this bar chart, each colour represents one type of edge, indicating different types of evidence for a relationship between two variables, and the width of a bar indicates how many of that type are present within each community.

communities with more *Hypothesis*-heavy edge profiles might hold more open questions, or perhaps it may suggest that communities such as *Early Language Evolution* have a higher potential for generating new scientific questions, whereas communities with fewer hypotheses might be more 'saturated' and less 'fertile' when it comes to new scientific ideas. We discuss this further, in the Discussion and Conclusion.

Fig. 11 shows the distribution of edges connecting different communities. Each pie chart describes the links between each pair of communities identified by the Louvain model (e.g. *Early Language Evolution* and *Sociolinguistics and Language Complexity*, also represented as *ELE* and *Soc*). The numbers underneath the pie charts indicate the number of links between each pair of communities, and the pie charts describe the proportion of different types of links, such as hypotheses, experiments, and simulations.

As shown in Fig. 11, the number of within-community links is considerably higher than the number of links across communities, which confirms that the Louvain method was indeed able to find a high-modularity partition of the network. When comparing communities, one also finds evidence that some communities act more as *sinks*, while others act more as *sources*: communities *Early Language Evolution*, *Linguistic Diversity and Geography*, and *Lexical Studies and Psycholinguistics* have more incoming links (i.e. a higher in-degree), indicating more hypotheses and other types of evidence that suggest that variables in those communities are influenced by variables in other communities, whereas communities *Cognition and Social Interactions* and *Neurolinguistics and Language Disorders* show the opposite pattern: their variables are more likely to be implicated in outgoing links, which more often indicate their effect on variables in other communities. Communities *Sociolinguistics and Language Complexity, Anthropological Linguistics and Kinship* and *Cultural and Ritualistic Elements* have a roughly equal in-degree and out-degree, suggesting a balance between incoming and outgoing links. This pattern is also reproduced when looking only at *Hypothesis* links (figure in the Appendix): some communities 'produce' a much higher number of hypotheses, while other communities 'receive' more hypotheses.

Finally, since betweenness centrality can also be defined for edges, we use that to sort the *Hypothesis*-type edges across different communities. This captures hypotheses which, if proven correct (or incorrect), would become part of the shortest path connecting large numbers of nodes that currently do not have many connections between them, thus bridging over large gaps in the network—which makes them potentially very fruitful avenues for future research. The top ten cross-community hypotheses, ranked according to their betweenness centrality, are presented in Table 2.

Discussion and conclusion

This study has traced the expansive trajectory of evolutionary linguistics from 1886 to 2022, revealing a field that has emerged from the combination of several disciplines into a cohesive, interdisciplinary, mature whole. The emergence of a giant connected component in the co-authorship network underscores the collaborative nature of contemporary research, bringing together diverse disciplines such as anthropology, neuroscience, and the social sciences. This interconnectedness is pivotal for the next steps of the discipline, and it is what we explore in this paper.

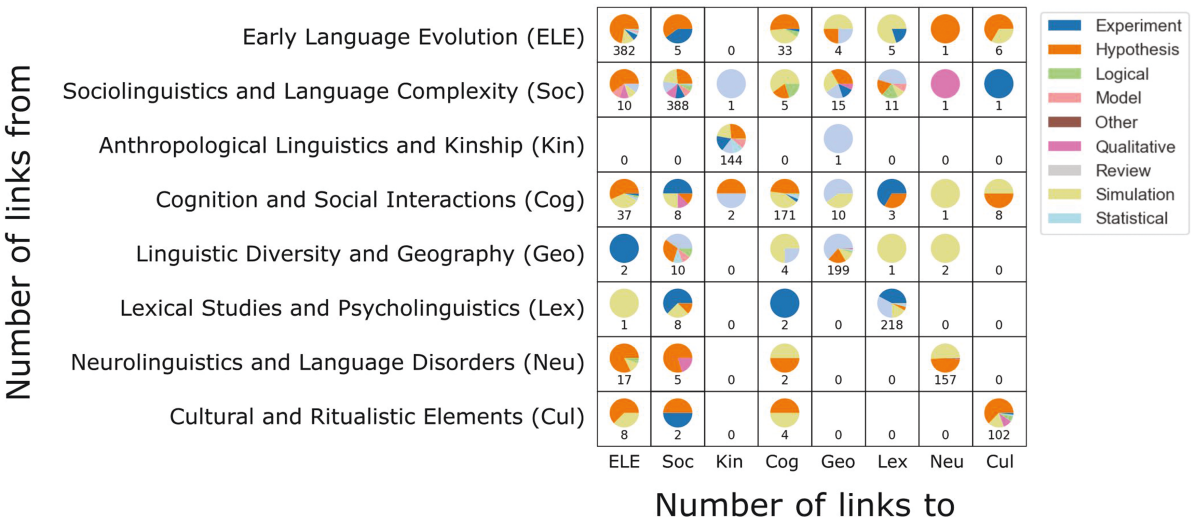

Figure 11 Each pie chart describes the links between each pair of communities identified by the Louvain model. Numbers underneath pie charts indicate the number of links between each pair of communities.

On the variable–variable network, we observe clusters of densely connected variables forming fan-like structures, typically occurring when a variable is hypothesized to have a causal effect on multiple other variables (or to be affected by them). These structures also tend to appear at the periphery of the network, indicating that those clusters of variables tend to be less central or important to evolutionary linguistics—and indeed, in CHIELD, those were typically genetics and neuroscience variables, which despite relevant to the evolution of language, are not so as central as *social interaction* or *population size*.

On the author-author network, we find clear evidence of how some authors act as bridges across communities of scientists. While those scientists are not necessarily the ones who publish the most or with most people, nor is their betweenness centrality unusually high for a network of this size (it is rather the opposite, values are very much within expected from the bootstrap sample), they hold an important role in the network, as they act as translators of knowledge between disciplinary communities. It's important to clarify that neither CHIELD nor this paper seeks to present a comprehensive or representative overview of the entire field of evolutionary linguistics scholarship. The vast and diverse nature of this field, encompassing various disciplines, methodologies, and research questions, makes such a comprehensive representation challenging. Moreover, being a crowdsourced dataset, CHIELD is subject to multiple biases. Most contributors to CHIELD are themselves academics in evolutionary linguistics, who will inevitably include papers they are familiar with to the point of being able to identify key variables and causal hypotheses connecting them. Those contributors are also typically in English-speaking developed countries, which further adds to the bias. Something could also be said about the type of paper being annotated: it is possible that even within the same community (e.g. *Early Language Evolution*), some papers will more likely be annotated by others—perhaps papers that are shorter or that have clearer writing and stating of their hypotheses, or even papers written by more prestigious authors.

Perhaps most importantly, a key limitation of any analysis relying on CHIELD is how variables within the database are coded on different levels of generality. While some are straightforward to operationalize and measure (e.g. 'obligatory grammatical distinction: future time reference' or 'population size'), others are less so (e.g. 'self-domestication'), and others—in particular 'language' and 'cooperation'—are extremely broad and polysemous concepts (Haspelmath 2020; Wacewicz et al. 2023). As such, the many edges featuring 'language' most likely refer to a broad family of partly or completely disjunct constructs corresponding to the different senses of this term—quite possibly constructs such as the capacity for language, or particular languages, or the entire domain of phenomena related to language use, as mentioned above (Haspelmath 2020). While this does not invalidate the results here presented, it does call for careful interpretation.

Rather than aiming for full coverage of the field, which would also require some disambiguation of terms such as 'language' or 'cooperation', the focus of this study has been more targeted, aiming to analyse the connections between authors and between variables

Table 2. The top ten cross-community hypotheses, ranked according to their edge betweenness centrality.

Hypothesis	Communities (to and from)	When stated
Brain size is correlated with population size	Neurolinguistics and Language Disorders, Sociolinguistics and Language Complexity	2016
Cooking is a sign of cooperation within human populations	Early Language Evolution, Cognition and Social Interactions	2013
A change in female reproductive strategy led to 'sham menstruation' signals	Cognition and Social Interactions, Cultural and Ritualistic Elements	1995
Protolanguage exerted evolutionary selection pressure on spoken social interactions	Early Language Evolution, Cognition and Social Interactions	2016
When a group moves into an already populated area, inter-group contact affects linguistic diversity	Linguistic Diversity and Geography, Sociolinguistics and Language Complexity	2013
Social bonding is a necessary precondition for larger group size	Early Language Evolution, Cognition and Social Interactions	2014
Social interaction causes more cooperation	Cognition and Social Interactions, Early Language Evolution	2007
Capacity for cooking led to residential changes	Cognition and Social Interactions, Anthropological Linguistics and Kinship	2014
Exchange of symbolic signals led to the emergence of language	Cultural and Ritualistic Elements, Early Language Evolution	2003
Compositionality does not causally influence protolanguage	Sociolinguistics and Language Complexity, Early Language Evolution	2007

within the field as represented in the CHIELD database. By examining the networks of causal hypotheses between variables and the networks of papers and authors, we have aimed to identify potentially significant areas for future research. It is worth considering that the lack of connection between subfields alone is not necessarily a problem nor an issue to be tackled: different scientific (sub)disciplines might be disconnected because they focus on different levels of explanation, or perhaps they are simply interested in different objects of study. This is another reason for our focus on edges that have already been proposed as hypotheses have not been empirically validated—and in particular, those edges or connections between variables of high centrality in the network, which might be impactful if proven right (or wrong), and perhaps ripe for further exploration.

In moving forward, it would be beneficial for future work to consider both CHIELD and the data collected and analysed by Wacewicz et al., among other resources, to build a more nuanced understanding of the field's collaborative networks. Such a comparison might highlight differences or similarities in the way research communities are structured and interact within the broader field of evolutionary linguistics. Wacewicz et al'.s analysis, which employed different methodologies and a different dataset, could serve as a valuable point of reference for understanding the dynamics of scholarly collaboration and the evolution of research themes over time.

There is another potentially interesting research avenue in comparing co-authorship networks, as the ones studied both in this paper and by Wacewicz et al., but also networks of authors which are linked whenever they have worked on the same variable, or set of variables (maybe even a community of them). This type of analysis would allow one to identify networks of authors with shared interests but who have not worked together, and might be valuable in studying how the scientific landscape of evolutionary linguistics is distributed. The overlap between these two types of author-author networks should be informative on how fragmented the field is: when these two networks are identical, this would suggest that variables (or groups of variables) are essentially under the monopoly of a group, in that working on variable X implies working with the people who work on variable X. Should the networks be fairly different, this result would suggest a more diverse and dispersed organization of the discipline.

In a broader context, this paper shows how the application of network science tools to crowdsourced databases such as CHIELD can produce very valuable insights for the study of a scientific discipline such as evolutionary linguistics. It also reveals many promising directions for future research: analyses similar to the ones developed here might be able to identify potential hypotheses which, if tested, would form a bridge between communities of variables disconnected until then, as was done in the Results section; or, in the spirit

of the original CHIELD paper, might help identify potential hypotheses that, when tested, should have downstream consequences, and help test competing theories. This will be particularly valuable to the evolution of language, but more broadly, to any sciences with multiple coexisting independent theories.

Finally, it is worth pointing out how our contribution is not limited to evolutionary linguistics, or even to broader cultural evolution studies. Rather, it is an approach that could be applied to any scientific database representing well-defined relationships between variables or concepts. Whether by identifying 'low-hanging fruit' research projects, or hypotheses with many downstream consequences, or even studying how the whole variable network evolves, we believe the approach we present here can make a significant difference to the understanding of the impact of the network of scientific collaborations on the production and collective construction of knowledge.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

Conflict of interest

We declare no conflicts of interest with this study.

Funding

No funding was received for conducting this study.

Data Availability

None declared.

References

Barabási, A-L. (2016) *Network Science*. Cambridge University Press, Cambridge, United Kingdom.

Bastian, M., Heymann, S., and Jacomy, M. (2009) 'Gephi: an open source software for exploring and manipulating networks'. Proceedings of the International AAAI Conference on Web and Social Media. Vol. 3. No. 1.

Bergmann, T., and Dale, R. (2016). 'A Scientometric Analysis of Evolang: Intersections and Authorships'. In: *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. New Orleans: Evolang Scientific Committee. http://evolang.org/neworleans/papers/182.html

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). 'Fast Unfolding of Communities in Large Networks', *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Cheney, D. L., and Seyfarth, R. M. (2005). 'Constraints and Preadaptations in the Earliest Stages of Language Evolution', *The Linguistic Review*, 22(2-4), 135–159. https://doi.org/10.1515/tlir.2005.22.2-4.135

Christiansen, M. H. (2003). 'Language Evolution: The Hardest Problem in Science?'. In: Morten H. Christiansen, and Simon Kirby (eds) *Language Evolution*. Oxford, United Kingdom, 2003. https://doi.org/10.1093/acprof:oso/9780199244843.003.0008

Croft, W. (2000). *Explaining Language Change: AnEevolutionary Approach*. Pearson Longman, London, United Kingdom.

Davidson, I. (2003) 'The Archaeological Evidence of Language Origins: States of Art'. In: Morten H. Christiansen, and Simon Kirby (eds) *Language Evolution*. Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780199244843.003.0008

Fortunato, S., Bergstrom, C. T., Börner, K., et al. (2018). 'Science of Science', *Science*, 359(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Frank, S., and Smith, K. (2020). 'Natural Population Growth can Cause Language Simplification', *EvoLang13 Proceedings*, 111, 14–111.

Gardner, R. C. (1983). 'Learning Another Language: A True Social Psychological Experiment', *Journal of Language and Social Psychology*, 2(2-3-4), 219–239. https://doi.org/10.1177/0261927x8300200209

Gong, T., Shuai, L., and Wu, Y. (2013). 'Multidisciplinary Approaches in Evolutionary Linguistics', *Language Sciences*, 37, 1–13. https://doi.org/10.1016/j.langsci.2012.09.002

Guevara, M. R., Hartmann, D., Aristarán, M., et al. (2016). 'The Research Space: Using Career Paths to Predict the Evolution of the Research Output of Individuals, Institutions, and Nations', *Scientometrics*, 109(3), 1695–1709. https://doi.org/10.1007/s11192-016-2125-9

Haspelmath, M. (2020). 'Human Linguisticality and the Building Blocks of Languages', *Frontiers in Psychology*, 10, 3056. https://doi.org/10.3389/fpsyg.2019.03056

Hauser, M. D., and W. Tecumseh Fitch. (2003) 'What Are the Uniquely Human Components of the Language Faculty?'. In: Morten H. Christiansen, and Simon Kirby (eds) *Language Evolution*. Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780199244843.003.0009

Hoff, E. (2006). 'How Social Contexts Support and Shape Language Development', *Developmental Review*, 26(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Hurford, J. R. (2003) 'The Language Mosaic and its Evolution'. In: Morten H. Christiansen, and Simon Kirby (eds) *Language Evolution*. Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780199244843.003.0003

Ke, J., and Holland, J. H. (2006). 'Language Origin from an Emergentist Perspective', *Applied Linguistics*, 27(4), 691–716. https://doi.org/10.1093/applin/aml033

Kirby, S., and Christiansen, M.H. eds. (2003). Language evolution. Oxford University Press, Oxford.

Lieberman, P. (2003) 'Motor Control, Speech, and the Evolution of Human Language'. In: Morten H. Christiansen, and Simon Kirby (eds) *Language Evolution*. Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780199244843.003.0014

MacWhinney, B. (2008). 'Cognitive Precursors to Language'. *The Evolution of Communicative Flexibility*, 1, 193–213. https://doi.org/10.7551/mitpress/9780262151214.003.0009

McMahon, A., and McMahon, R. (2012). *Evolutionary Linguistics* (Vol. 223). Cambridge University Press.

Newman, M. (2010). *Networks*. Oxford University Press, Oxford, United Kingdom.

Newman, M. E. (2006). 'Modularity and Community Structure in Networks', *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nölle, J., Hartmann, S., and Tinits, P. (2020). 'Language Evolution Research in the Year 2020: A Survey of New Directions', *Language Dynamics and Change*, 10(1), 3–26. https://doi.org/10.1163/22105832-bja10005

Peixoto, T. P. (2023). *Descriptive vs. Inferential Community Detection in Networks: Pitfalls, Myths and Half-Truths*. Cambridge University Press, Cambridge, United Kingdom.

Roberts, S. G., Killin, A., Deb, A., et al. (2020). 'CHIELD: The Causal Hypotheses in Evolutionary Linguistics Database', *Journal of Language Evolution*, 5(2), 101–120. https://doi.org/10.1093/jole/lzaa001

Scott-Phillips, T. C., and Kirby, S. (2010). 'Language Evolution in the Laboratory', *Trends in Cognitive Sciences*, 14(9), 411–417. https://doi.org/10.1016/j.tics.2010.06.006

Számadó, S., and Szathmáry, E. (2006). 'Selective Scenarios for the Emergence of Natural Language', *Trends in Ecology & Evolution*, 21(10), 555–561. https://doi.org/10.1016/j.tree.2006.06.021

Tattersall, I. (2014). 'An Evolutionary Context for the Emergence of Language', *Language Sciences*, 46, 199–206. https://doi.org/10.1016/j.langsci.2014.06.011

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, Massachussets, USA.

Wacewicz, S., Sibierska, M., Placiński, M., Szczepańska, A., Poniewierska, A., Ng, Y., and Żywiczyński, P. (2023). 'The Scientometric Landscape of Evolang: A Comprehensive Database of the Evolang Conference', *Journal of Language Evolution*, 7.2, lzad003.