A SCIENTOMETRIC ANALYSIS OF EVOLANG: INTERSECTIONS AND AUTHORSHIPS

TILL BERGMANN[1] & RICK DALE[1]
[1]*Cognitive and Information Sciences*
University of California, Merced
Merced, CA 95340, USA

Research on the evolution of language has grown rapidly, and is now a large and diverse field. Because of this growing complexity as a scientific domain, seeking new methods for exploring the field itself may help synthesize knowledge, compare theories, and identify conceptual intersections. Using computational methods, we analyze the scientific content presented at EvoLang conferences. Drawing on 365 abstracts, publication patterns are quantified using Latent Dirichlet Allocation (LDA), which extracts a semantic summary from individual abstracts. We then cluster these semantic summaries to reveal the frameworks and different domains present at EvoLang. Of course, our results show that EvoLang is an interdisciplinary field, attracting research from various fields such as linguistics and animal studies. Furthermore, we show that the framework of iterated learning and cultural evolution is a hub topic at EvoLang.

1. Introduction

In this paper, we explore the conceptual structure of research on language evolution itself by analyzing the submissions to the EvoLang conference over the past 10 years. Our goal is to provide insight into the network of theories, concepts, and methods that populate this growing field. Since its inception in 1996, EvoLang has become a prominent and well-attended conference. It is now the premiere conference on language evolution, with more than 100 presentations at the last EvoLang in Vienna and over 300 delegates in attendance. This is a five-fold increase from the first EvoLang in 1996. How might we quantify this rapidly growing scientific content?

There are numerous reviews of language evolution which attempt to unpack and relate its various theories and debates (e.g. Christiansen & Kirby, 2003; Bickerton, 2007; Fitch, 2010). These provide impressive coverage, especially considering the diversity and complexity of language evolution research. Research at EvoLang tackles a wide range of these topics, spanning the many levels of language, from the evolution of flexible signalling strategies, to the social cognitive processes that may undergird human linguistic skills.

In what follows, we use topic modeling (Griffiths & Steyvers, 2004; Yau, Porter, Newman, & Suominen, 2014) to extract the set of latent conceptual topics

that make up EvoLang. We find that there are three distinct conceptual clusters that can be inferred from the abstracts, including the iterated learning framework and comparative studies. Second, we combine these topic clusters with a co-authorship network analysis to assess the relative influence of these typical topic clusters, finding that the iterated learning cluster in particular serves as a central hub in the broader EvoLang community. By analyzing the knowledge bases of EvoLang, it may be possible to attain a firmer grip on the state of the art in the field, and the relationships among its various theories.

2. Modeling the content of EvoLang submissions

We selected all abstracts from submissions between 2006 and 2014 with more than 500 characters.[a] We then applied a Latent Dirichlet Allocation algorithm (Blei, Ng, & Jordan, 2003) on the resulting 375 abstracts, a method that is commonly used in scientific content analysis (Griffiths & Steyvers, 2004). In LDA, each document (here, abstract) is represented by a distribution over topics, and the topics themselves are represented by a distribution over words. That is, each topic consists of a distribution of semantically related words, and each abstract can then be represented as a combination of these topics, which make up the *gist* of the document. For example, one abstract at EvoLang may combine the topics of non-human communication and learning, while another may combine syntax and computation. Importantly, the algorithm first extracts numerically identified topics, which are then interpreted by the researcher – so the example topic combinations here are simply hypothetical. Researchers typically inspect numbered topics (topic 1 ... topic k), and these topics are then interpreted from their associated words. As we show below, this can result in a compelling intuitive set of topics.

After running the algorithm with a various number of topics, we selected the model of best fit, which contained 20 topics. Example topics are shown in Table 1 with associated terms. Note in the table that we have used a stemmer algorithm to obtain roots (e.g., "compar", "abil"), to decrease the type-token ratio, and facilitate topic extraction. To further analyze the content, a correlation matrix of the probability distributions for the topics was calculated and a network of positively related topics was generated. Then, a community detection algorithm (Pons & Latapy, 2005) was used to cluster these topics. We found that the algorithm clustered the content of EvoLang submissions broadly into three communities or clusters. The resulting network is shown in Fig. 1, with the different clusters marked by color.

But what do these clusters consist of? To get more insight into the topics associated with each cluster, we extract the most probable terms associated with the

[a] Abstracts before 2006 were published in a different format and were thus omitted to keep the data consistent.

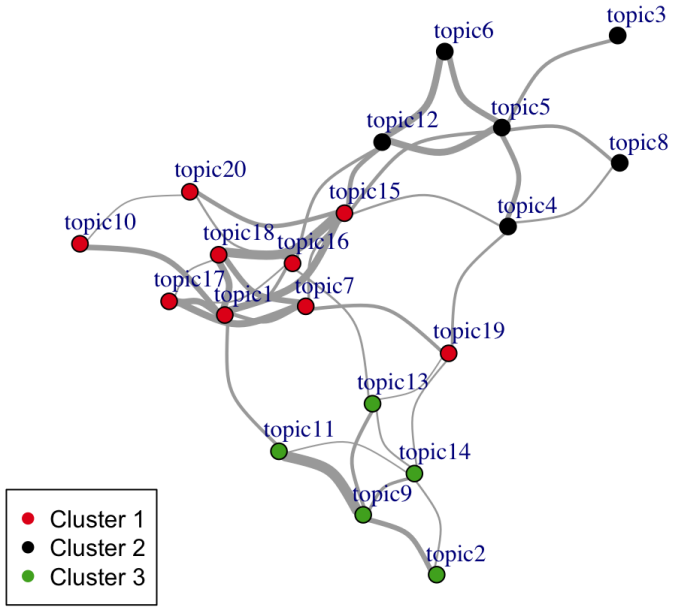

Figure 1. Network of positively correlated topics. The thicker an edge, the stronger the correlation. Topics belonging to the same cluster share a color. An interactive version of this plot is available on http://shiny.tillbergmann.com/apps/evolang/

topics in each cluster. The first cluster covers general terms covering linguistics and language evolution, but also more specific topics such as word order in topic 19 (see Table 1). The second cluster is more specific, including comparative studies involving primates and birds, as well as the study of gestures and music (Table 2). Lastly, papers in the third cluster approach language evolution through cultural evolution and iterated learning, as well as the emergence of structures in communication experiments (Table 3). Inspecting these terms and communities gives a good overview of different fields within EvoLang, and indeed, both the clustering and most probable terms make intuitive sense.

In general, these clusters show that EvoLang hosts a variety of sub-fields,

which approach the study of language evolution from varying angles. Not only does it include more theoretical linguistic work, but also comparative studies are well represented. Certainly this is well known intuitively by researchers within the community, but the analysis here suggests that there are crisp clusters that can be automatically extracted using the topic model. In the next section, we look at the author collaboration networks of EvoLang. This serves both as an illustration of the range of authorship patterns, as well as being the measure through which we further analyze the interconnectedness of these three topic clusters.

Table 1. Terms associated with cluster 1.

Topic 1	Topic 7	Topic 10	Topic 15	Topic 16
languag	semant	evolut	human	symbol
evolut	evolutionari	select	abil	evolutionari
evolv	grammar	extend	language	icon
evolution	syntax	behavior	research	language
language	structur	term	share	protolanguag
framework	approach	factor	compar	sound

Topic 17	Topic 18	Topic 19	Topic 20
system	process	word	signal
evolut	cognit	order	communic
paper	brain	inform	mechan
complex	evolut	divers	behaviour
increas	specif	cue	explain
stage	propos	speaker	provid

Table 2. Terms associated with cluster 2.

Topic 3	Topic 4	Topic 5	Topic 6	Topic 8	Topic 12
modern	question	vocal	speech	song	gestur
present	differ	human	origin	learn	communic
suggest	music	primat	function	development	ape
air	speech	call	involv	genet	studi
evid	pattern	product	action	finch	intent
homo	show	produc	area	complex	system

Table 3. Terms associated with cluster 3.

Topic 2	Topic 9	Topic 11	Topic 13	Topic 14
communic	learn	linguist	emerg	model
studi	mean	cultur	languag	social
game	categori	bias	develop	agent
refer	experi	evolut	form	popul
experiment	structur	languag	sign	network
strategi	iter	learn	languages	interact

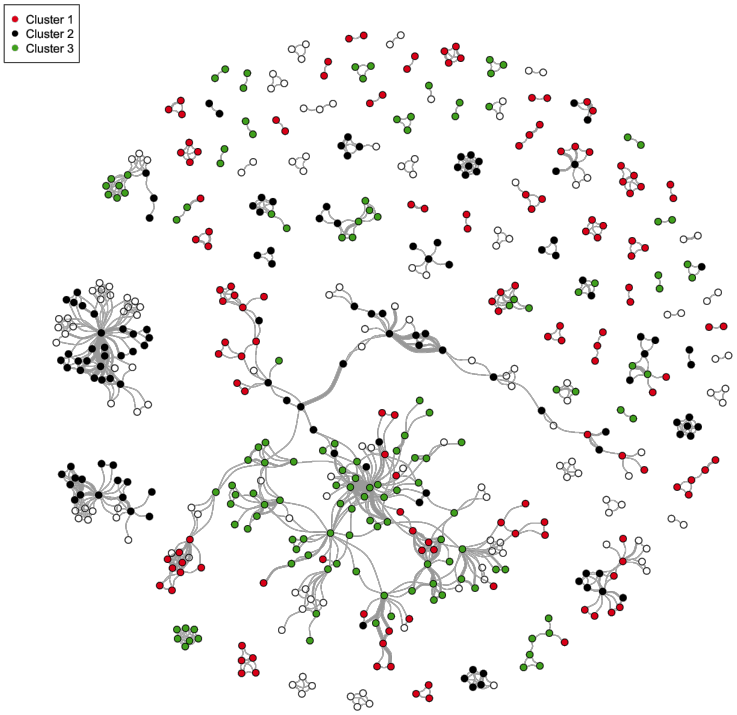

Figure 2. A network showing collaborations between authors. Nodes represent authors and are colored with respect to their dominant cluster. The thicker an edge, the more collaborations between the nodes. An interactive version of this plot is available on http://shiny.tillbergmann.com/apps/evolang/.

3. The interconnectedness of authors and clusters

By constructing an authorship network from co-authored abstracts, we can detect which authors have a high interconnectedness at EvoLang. Authors who publish and collaborate often are referred to as "central," and by virtue of their centrality, we can also assess the contribution of their associated topics in their collaborations. In this network, each node is an author, and each edge between two nodes represents collaboration between these two nodes/authors. Edge weight (connection strength) is determined by the number of collaborations between these two authors. Using the topic clusters from the above analysis, we calculated the most prevalent cluster for each author, based on which cluster their respective papers were assigned. By plotting the author network (Fig. 2), we can see that there are some hubs in the middle of the network, as well as some collaborations outside these general hubs, not connected to the rest of the network. These smaller

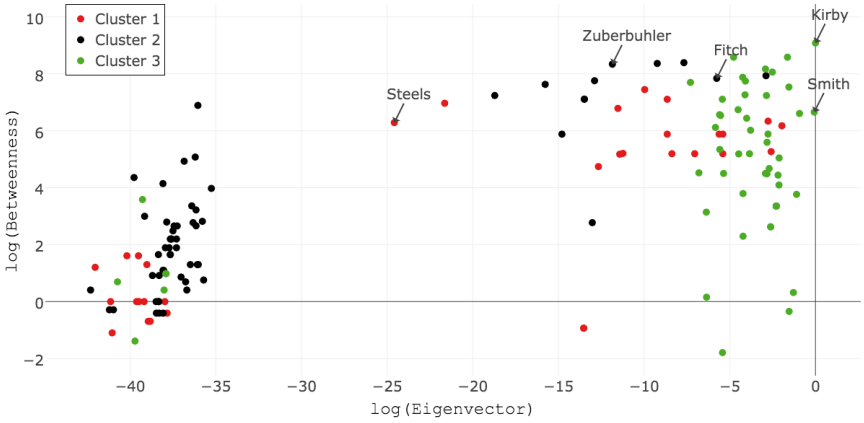

Figure 3. Betweenness and eigenvector centrality, on a log-scale. Each point represents an author, with the color representing their cluster. A few noteworthy authors are labeled. An interactive version of this plot is available on http://shiny.tillbergmann.com/apps/evolang/.

collaborations often consist of advisor-advisee relationships within the same lab or department. The color of the nodes represents the respective cluster an author has mainly published in. As not all papers were included in the content analysis due to abstract length, some nodes remain white because their cluster could not be determined. The bigger hubs in the center of the network mainly belong to cluster three, covering the iterated learning framework. Cluster 1 and 2 are more interspersed, and cluster 2 forms its own smaller hubs, showing a strong sense of collaboration in comparative studies.

After constructing the network, centrality measures were used to detect the most influential authors within this network. In network theory, there are multiple ways to measure the centrality of nodes (Freeman, 1978; Koschützki et al., 2005). Here, we look at two values: eigenvector centrality and betweenness centrality. Eigenvector centrality measures the influence of a node by assigning a score based on connections to high scoring nodes (here, nodes with a lot of collaborations and thus submitted papers). The score is bound between 0 and 1, with 1 representing highest centrality. Betweenness centrality assigns a score based on how often the node is part of the shortest path between two other nodes, and thus measures how well a node connects different parts of a network. These nodes are considered to be important in communication between other nodes and keeping the network connected. Fig. 3 shows the centrality measures of authors on a log-scale (purely for illustrating purposes): Authors with high eigenvector values but low betweenness have close contact to important people, while authors with low eigenvector values but high betweenness values serve as valuable connections between nodes.

In the plot, there is a division between authors with a high and low Eigenvector centrality. Authors with a high Eigenvector centrality tend to be in cluster 3, while authors in cluster 1 are more likely to have low Eigenvector centrality. Cluster 2 authors seem to be more interspersed.

Table 4. Summary statistics for each cluster of topics.

Cluster	M(Eigenvector)	SD(Eigenvector)	M(Betweennes)	SD(Betweenness)
1	0.003033	0.01517	63.29	227.6
2	0.002757	0.01482	220.42	782.2
3	0.037867	0.12790	336.23	1113.3

Table 5. Summary of multinomial logistic regression showing log-odds and standard errors.

	Dependent variable:	
	Cluster 1	Cluster 2
Betweenness	−0.001	0.0001
	(0.0003)	(0.0002)
Eigenvalue	−21.989*	−25.957*
	(0.001)	(0.001)
Constant	0.275*	0.137
	(0.128)	(0.132)
Akaike Inf. Crit.	866.008	866.008

By using the centrality measures calculated for each author, we were able to deduce the influence of each topic cluster. That is, to which cluster do the most widely collaborating individuals belong? Table 4 shows summary statistics for the author centrality measures in each cluster. Not surprisingly, cluster 3 has both the highest average eigenvector and betweenness centrality, however, it also has the highest deviations. While the deviations suggest that there is a lot of variation within clusters, it looks like cluster 3 is the most central set of topics within EvoLang.

To test whether this difference in centrality measures is significant, a multinomial logistic regression was run with the clusters as a dependent variable, and the two centrality measures as the independent measures. Cluster 3 was chosen as the baseline community, as we hypothesized that it had higher centrality than the other two clusters. The model output is summarized in Table 5 and was significant compared to a null model ($\chi^2(4) = 44.208, p < 0.0001$). Significance values were calculated using Wald tests. Coefficients for betweenness centrality were not significant (Cluster 1: $p = 0.08$, Cluster 2: $p = 0.59$). However, eigenvector centrality was a significant predictor for both cluster ($p < 0.0001$ for both

clusters). As the log odds are very high, any increase in eigenvector centrality increases the probability of that a paper is in cluster 3.

From this analysis, we conclude that cluster 3, which appears strongly related to iterated learning and cultural evolution, serves as a "hub cluster" within EvoLang. However, as the betweenness centrality was not a significant predictor of cluster/framework, authors within each cluster serve as an important connection between other authors, and clusters as a whole.

4. Summary

We analyzed the content of abstracts presented at EvoLang. Our analysis of latent topics shows that EvoLang is an interdisciplinary conference, and seems to draw from three major clusters of topics. Using a network analysis of author collaborations, we investigated these clusters with regard to their influence. Our results suggest that the iterated learning and cultural evolution framework is associated with a high centrality property within EvoLang. Comparative studies with primates are an important interconnector between authors and communities, while the cluster covering linguistic approaches is interspersed and well represented throughout the conference. The interconnectedness of the author network suggests that each cluster draws inspiration from each other, and that in fact no single framework – according to the LDA topic model – is isolated from any other.

Though these patterns may be intuitive to highly initiated attendees of the conference, the purpose of this paper is to demonstrate that scientometric techniques can be used to reveal these patterns quantitatively. With just under 400 abstracts, a number of natural authorship and conceptual patterns emerge. It may be useful and interesting to carry out similar analyses in subsequent years to discern how this field is changing, and how topic clusters may be converging or co-fertilizing.

References

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua*, *117*(3), 510–526.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Christiansen, M. H., & Kirby, S. (2003). Language Evolution: The Hardest Problem in Science? In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 1–15). Oxford: Oxford University Press.

Fitch, T. W. (2010). *The Evolution of Language*. Cambridge, MA: Cambridge University Press.

Freeman, L. C. (1978). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, *1*(3), 215–239.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, *101*, 5228–5235.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality Indices. In U. Brandes & T. Erlebach (Eds.), *Network Analysis* (pp. 16–61). Berlin and Heidelberg: Springer.

Pons, P., & Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In P. Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and Information Sciences - ISCIS 2005* (Vol. 3733, p. 284-293). Berlin and Heidelberg: Springer.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786.