User Behaviour as a Predictor of Grammatical Similarity within Digital Networks

Keywords: Language, Social Media, Emergence, Complex Adaptive Networks

Extended Abstract

Motivation. The availability of social media language data, such as from Reddit, offer us a unique opportunity to analyse a linguistic system in its entirety. More importantly, this enables us to observe variation and change within a complex adaptive system (CAS). Dialectologists frequently rely on the distribution of discrete variables. As an example, 'in the weekend' in New Zealand English is preferred to 'on the weekend' in American English in contrast to 'at the weekend' in older varieties of British English [3]. However, this top-down approach does not account for the interactions that exist within a linguistic system. Computational Construction Grammar (C2xG) takes a bottom-up approach to 'learn' a grammar based on lexical, syntactic, and semantic constraints [2]. Unlike large language models (LLMs), C2xG extracts interpretable grammatical constructions within a linguistic system. We therefore ask, can we predict grammatical similarity based on user behaviour within a digital network.

Approach and Methodology. We focus on the linguistic context of New Zealand Reddit. Our primary data set included 34 Reddit communities associated with New Zealand (744 million words) from *Pushshift* [1]. In order to streamline our analysis, we limit our corpus to place-based Reddit communities associated with the six major cities of New Zealand. We hypothesise that user behaviour predicts grammatical similarity. Our methodological pipeline included three phases. **Phase 1) Model Development**: we first trained a C2xG from the New Zealand-related Reddit communities. The final grammar consisted of 1,582 distinct lexical (LEX-Only) features, 1,026 syntactic (SYN-Only) features, and 8,284 semantic (SEM+) features. **Phase 2) Feature Extraction**: next, we grouped the language data by geography and user behaviour before parsing the city-level Reddit communities with our custom C2xG model. The two user behaviour variables included lifespan cohort and engagement ratio defined as the ratio of post and comment submissions to upvotes received. **Phase 3) Analysis**: once parsed, we then calculate cosine similarity as a measure of grammatical distance between geography and the user behaviour groupings. We use Ordinary Least Squares (OLS) Regression to evaluate the predictive ability of our dependent variable (grammatical similarity) and three independent variables (mean lifespan, upvote score, and user engagement).

Results. We focus on SYN-Only in our results. The results from our optimal OLS regression models for age cohort ($R2 = 0.662$; Adjusted $R2 = 0.645$; $F = 39.829$) and engagement ratio ($R2 = 0.668$; Adjusted $R2 = 0.651$; $F = 40.827$) suggest that as user lifespan and engagement increase within a city-level community. These findings are further supported by the similarity network structure (as shown in Figure 1). With the SYN-Only features, we constructed a similarity network to show how user cohorts based on engagement ratio (nodes) self-organise in relation to grammatical similarity (edge weights). Similar results were observed with the LEX-Only features.

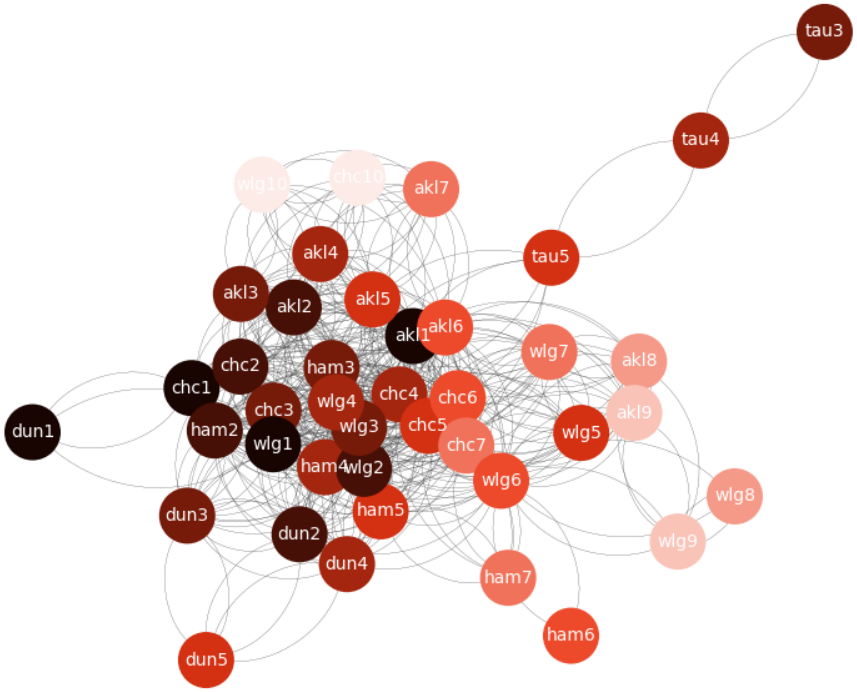

Figure 1: Network graph visualising the cosine similarity (greater than 99%) between city-level communities for syntactic features by engagement ratio. Each node represents a user grouping by engagement ratio (as deciles). The 'most engaged' user decile in the darkest shade of red (1), and the 'least engaged' user decile in the lightest shade of red (10).

Conclusions and Outlook. The results from our OLS regression models support our primary hypothesis that user behaviour was a predictor of grammatical similarity. Younger and least engaged users were the least grammatically similar, which suggests these user groups were more likely to be drivers of variation and change within this CAS. Further work is required to consider how similar predictive patterns persist across dialect and language conditions within digital networks, with the potential for spoken language.

References

[1] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839, Atlanta, GA, May 2020. PKP Publishing Services Network. https://doi.org/10.1609/icwsm.v14i1.7347.

[2] Jonathan Dunn. *Computational Construction Grammar: A Usage-Based Approach*. Cambridge University Press, 2024. https://doi.org/10.1017/9781009233743.

[3] Peter Trudgill and Jean Hannah. *International English: A Guide to Varieties of English Around the World*. Routledge, London, England, 6 edition, April 2017. https://doi.org/10.4324/9781315192932.