Newcastle University

Open Research Case Study

Translating Openness: Bringing Open Research Practice into Theoretical Linguistics

Xinyu Zhu

PhD student in Linguistics, School of English Literature, Language and Linguistics .

Introduction and Research context

As a third-year PhD student in Linguistics, I work within theoretical and descriptive traditions, studying the syntax of reduplicated adverbials in Mandarin Chinese. In these fields, research is often built upon data drawn from previous literature or the researcher's own subjective judgments. This long-standing practice, while deeply rooted in humanities, has been increasingly challenged for its lack of transparency, replicability, and data consistency. I came to realize, however, that openness in theoretical linguistics does not have to be limited to experiments or quantitative datasets. It can also mean making linguistic reasoning, data selection, and analytic procedures visible and shareable. Motivated by this idea, I began constructing a corpus-based dataset of Mandarin reduplicated adverbials, not as a purely empirical project, but as a way to provide objective, transparent, and reproducible evidence for theoretical claims that have traditionally relied on researchers' subjective analyses.

This project was also inspired by one of my supervisors, Dr. Cong Zhang, a 2024 Open Research Award winner, who introduced me to the concept of open research and encouraged me to apply it to theoretical inquiry. The resulting dataset brings an empirical and transparent dimension to linguistic theory, showing how theoretical and descriptive work can benefit from open, data-informed methods. It has been uploaded to GitHub and the Open Science Framework (OSF) and will be made fully public once the corresponding article completes peer review. This initiative aims to demonstrate that open science can find a home even in the most interpretive corners of the humanities

Open research practice

First, I took the initiative to learn programming languages such as Python and analytical tools like R, skills that are rarely part of formal training in the humanities. For me, learning them was not a technical exercise but a way to make my research process more open, traceable, and shareable. These tools enabled me to document data collection and analysis systematically so that others could follow the same steps. Next, I reflected on which aspects of my research would be most suitable and valuable, if shared openly, data that could benefit not only my own work but the broader field of theoretical and descriptive linguistics. Finally, I developed a semiautomated workflow that combined computational extraction with manual verification to build a dedicated dataset of Mandarin reduplicated adverbials, designing Python scripts to extract, organise, and verify the data. All scripts, dataset sources, instructions, and the resulting dataset were made publicly accessible on GitHub and OSF

Benefits

Making the dataset and workflow openly available has benefits that extend well beyond my own research. While the dataset itself is valuable, the greater contribution lies in the shared workflow and scripts, which record not only what was done but how it was done, a difference akin to giving someone a fish versus teaching them how to fish.

For linguists, this open workflow shows how interpretive analysis can be made systematic without losing its humanistic depth. It allows others to trace decisions, adapt the methods to new corpora, or question underlying assumptions. For students and early-career researchers, it offers a practical entry point into open research, demonstrating that computational tools can enrich rather than replace interpretation.

More broadly, the project helps foster a growing culture of openness in the humanities, showing that transparency and accessibility are achievable even in disciplines built on meaning and reflection.

Challenges, solutions and lessons learnt

Open research is not merely an attitude, it also requires technical skills and a carefully refined workflow. The first challenge was the learning curve. As a researcher trained in theoretical linguistics, I was unfamiliar with Python and digital tools essential for open practices. I addressed this by keeping an open mindset and learning gradually, viewing technical growth as part of becoming an open researcher.

The second challenge was deciding what information should be included and shared. I resolved this by reflecting on how my work could serve the wider community, aiming to create an openaccess resource that supports theoretical inquiry rather than technical perfection.

This experience taught me that open research in the humanities depends on both reflective awareness and practical skill. It means making interpretive decisions visible, documenting processes carefully, and offering a replicable model that others can build upon.

Conclusion

Open science in the humanities is not a barrier or an irrelevant methodology. I hope that my open research practice can serve as a useful reference for other researchers in theoretical linguistics and across the broader humanities.