Newcastle University

Open Research Case Study

Open science practices in linguistics research

Cong Zhang

Lecturer in Phonetics and Phonology at the School of Education

Introduction and research context

I am a Lecturer in Phonetics and Phonology at the School of Education, Communication and Language Sciences. My research primarily focuses on speech prosody, speech technology, and linguistics data collection methods. While linguistics is traditionally regarded as a humanities discipline, many subfields have moved towards a more empirical direction that calls for openness and reproducibility. Therefore, I adopt open science as the standard practice for all my research and advocate for its broader adoption within the field. In my studies involving model training and applications, my collaborators and I ensure that all source codes and training data are openly accessible (Case 1). For studies with data analysis, the data and analysis scripts are shared (Case 2). Even for studies without data, I make publications publicly available through preprint services, e.g. arXiv and OSF. Furthermore, I actively participate in initiatives advocating for changes in research culture to further support open science (Case 3).

Open research practices

Case 1: Open-source models & tools

Colleagues and I have developed tools such as *Charsiu Aligner*, *CharsiuG2P*, and *rhythm.metrics*. All associated training data, scripts, and publications are made available online. When appropriate, we also upload the models to *Hugging Face*, a platform for open-source AI models. Notably, for *CharsiuG2P* – a tool that converts writing to sound transcriptions – we also collated open-source datasets from 100 languages, many of which are low-resource languages. We standardised the formats of those data and included comprehensive metadata, including data licenses.

Case 2: Open-source datasets & analysis scripts

During the COVID-19 pandemic, in-person data collection was extremely difficult, impacting phonetics data collection which relies on high-quality recordings for data analysis. To address this challenge, my collaborators and I collected recordings simultaneously made using phones, computers, and lab-quality recorders. Recognising the difficulty of acquiring such data during the pandemic, we shared these datasets and analysis scripts [1, 2, 3], aiming to support other researchers in replication or further analysis. Additionally, we recently published a protocol for making high-quality recordings remotely, which can further support data collection in research and teaching.

Case 3: Changing research cultures

I am part of the leadership team for the *ManyLanguages* Initiative, which promotes open science by addressing linguistics questions through collaborative efforts across different languages. Such *big team science* projects inherently require openness, with all resources – including protocols, methodologies, and data – shared among researchers. We plan to facilitate a few projects to foster a culture that embraces more open-science big-team projects.

Additionally, I have been working on promoting gamified data collection through citizen science, partly funded by the Pioneer Award, Institute of Social Science. We have developed a game prototype for data collection. The gamified approach not only accelerates data collection but also engages the public in scientific research. A recently published position paper demonstrates how gamified data collection works in applied linguistics.

Benefits

The above three cases illustrate how the two key drives for open science are realised in my research:

(1) *quality reassurance* – *Case 1* and *2* ensured that all resources for reproducible science are transparent and accessible to other researchers, which can enhance the quality of scientific research.

(2) *knowledge sharing* – *Case 3* is a step forward towards open science by challenging the status quo, and further promotes knowledge sharing between researchers. The gamified approach further extends inclusivity and impact towards the general public.

Challenges, Solutions and Issues

Making science open is time-consuming. Organising resources for public scrutiny involves extra steps, such as annotating data analysis scripts and responding to detailed inquiries about open-source tools. Some challenges can be mitigated by documenting detailed instructions during development, but many lack easy solutions. Nonetheless, I view these efforts as essential to high-quality research and encourage their adoption.

Shifting research culture is even more challenging. Under the current research culture, big-team projects are uncommon and face systemic hurdles, such as the undervaluation of authorship in a 100-author paper compared to a single-author one. This is not an issue that can be resolved overnight but we have been working first towards normalising them. We will continue to nudge the boundaries a little at a time.

Conclusion

Incorporating open science practices into linguistics research has significantly enhanced the transparency, reproducibility, and accessibility of my work. Despite the challenges, the benefits to both the research community and the public are substantial. I remain committed to advocating for and implementing these practices, striving to foster a more inclusive and collaborative research environment.