National Spoken Language Database Committee

Report to the ASSTA Annual General Meeting on 4 December 2024

The National Spoken Language Database (NSLD) Committee is responsible for the promotion of Australian spoken language data collection and for the maintenance of two speech data corpora, the *Australian National Database of Spoken Language* (ANDOSL) and the audiovisual Australian speech corpus *AusTalk*. For safekeeping and to ensure their enduring accessibility and usability for research purposes, both corpora were transferred to ASSTA by the researchers who collected the data.

The NSLD Committee has comprised Steven Bird (Charles Darwin University), Steve Cassidy (Macquarie University), Jonathan Harrington (Ludwig Maximilian University, Munich), Julie Vonwiller (Founder of Appen Ltd) and Michael Wagner (NCBS Pty Ltd) as Chair. The Committee's, particularly Steve Cassidy's, activity over the last 12 months has been focused on making the audio part of the AusTalk corpus available through *Language Data Commons of Australia* (LDaCA). One copy of the video data is held in the Queensland Cyber Infrastructure Foundation (QCIF) archive, and a partial copy is apparently held by Prof. Roberto Togneri at the University of Western Australia. AusTalk is currently not readily accessible, and it will be one of the task of the future NSLD Committee to make the audio part accessible and usable and to explore pathways to achieve accessibility and usability for the video part of the corpus.

All five Committee Members have decided to step down effective 31 December 2024 to make way for a new generation of Members. At the same time, they expressed unanimously the wish that the two corpora should remain accessible in the case of ANDOSL and become accessible in the case of AusTalk and that a new NSLD Committee should again take charge of that task as well as promoting the collection of new Australian spoken language data in the future.

The Committee is conscious of the fact that the requirements for ensuring the continued accessibility and usability differ considerably between the two corpora. ANDOSL is well documented and quite well structured with a size of just under 10GB and could be able to be accommodated on the ASSTA server without major problems. AusTalk, on the other hand, has 3.7TB of audio data, is not yet as well-structured, and has no working interface currently to make the audio accessible on a server. An even larger problem is presented by the approximately 25TB of AusTalk video data in terms of both finding an appropriate host site and a suitable interface that is capable of making a database of that size accessible and usable. In addition, there is apparently an unresolved legal impediment of publishing participants' images without their explicit consent, which was first raised shortly after the AusTalk data collection had been completed.

Finally, the outgoing NSLD Committee wishes to make the following recommendations:

- We express our appreciation of the committed and expertly contributions over many years by Committee Member Steve Cassidy to making the AusTalk corpus accessible to the research community and particularly that his contributions were made in addition to his main job as a full-time academic.
- We recommend that ASSTA consider employing technical assistance, if not full-time then at least part-time, for the immediate tasks of making ANDOSL available through the ASSTA website or through a link from the ASSTA website and of making the AusTalk audio data available to the research community in the short term.
- We recommend that the incoming NSLD Committee continue to explore pathways of making the AusTalk video data available to the research community in the longer term.

(A potentially useful pathway is suggested by Julie Vonwiller: “In regard to getting the data into a searchable and accessible form, the incoming Committee could get advice from a business which also has enormous amounts of data for storage and use. Google Sydney may consider it a social contribution to help ASSTA do that; or CSIRO, the Australian Museum, the Powerhouse Museum (they have just almost completed digitising all their data), or Atlassian Sydney.”

Outgoing Committee Members, Steve Cassidy and Michael Wagner, are willing to help facilitate, as much as possible, the handover to the incoming NSLD Committee.

Prof. Michael Wagner,
Chair, National Spoken Language Database Committee