The LnNor Corpus: A spoken multilingual corpus of non-native and native Norwegian, English and Polish

Wrembel Magdalena, Hwaszcz Krzysztof, Pludra Agnieszka, Skałba Anna, Weckwerth Jarosław, Malarski Kamil, Cal Zuzanna, Kędzierska Hanna, Czarnecki-Verner Tristan, Balas Anna, Kaźmierski Kamil, Żychliński Sylwiusz, Gruszecka Justyna

The LnNor corpus was created as part of the data collection in two projects: CLIMAD (Cross-linguistic influence in multilingualism across domains: phonology and syntax) and ADIM (Across-domain Investigations in Multilingualism: Modeling L3 Acquisition in Diverse Settings), led by Prof. Magdalena Wrembel at Adam Mickiewicz University in Poznań, Poland and by Prof. Marit Westergaard at the Arctic University of Norway, from December 2021 to April 2024 with funding from the National Science Centre (NCN) in Poland and Norway Grants.

The CLIMAD and ADIM projects explored cross-linguistic influence (CLI) in the acquisition, processing, and use of a third language (L3/L*n*) across various language domains and focused on different settings and stages of acquisition from a multilingual perspective. A range of sophisticated methodologies, such as perception and production tests, grammaticality judgement tasks and online brain imaging techniques like EEG, were leveraged to unravel the intricacies of multilingual processing. By capturing real-time insights into the interplay of cross-linguistic influences, the projects not only provided valuable contributions to the understanding of L3/L*n* acquisition but also advanced theoretical frameworks in this field.

Corpus data collection covered a broad range of speech elicitation tasks. The recordings consist of word, sentence and text reading, picture story description, video story retelling, spontaneous speech and socio-phonetic interviews in Polish, English and Norwegian. The corpus contains metadata based on the Language History Questionnaire (Li et al. 2020) such as age, gender, native languages, proficiency level, length of language exposure, age of onset.

Data was collected from different **groups of speakers**:

- L1 Polish learners of Norwegian as L3/Ln, attending Scandinavian studies at Poznań College of Modern Languages and the University of Szczecin (instructed learners)
- L1 Polish learners of Norwegian as L3/Ln, living in Norway (naturalistic learners)
- L1 English natives as controls
- L1 Norwegian natives as controls
- speakers of L2/L3/L*n* English and L2/L3/L*n* Norwegian with various L1 backgrounds

Seven types of **speech tasks** were recorded in Norwegian, English and Polish:

- word reading
- sentence reading
- text reading (“The North Wind and the Sun”)
- picture description
- story telling
- video description
- translation from Polish/English into Norwegian

Metadata corresponding to the recordings include the following information:

- speaker ID, age, gender, education, current residence, speaker status (instructed/naturalistic/native), native language, additional languages spoken
- recording ID
- language: PL (Polish), EN (English), NO (Norwegian)
- status: L1, L2, L3/L*n*
- speech task: WR (word reading), SR (sentence reading), TR (text reading), PD (picture description), ST (story telling), VD (video description), translation from Polish (TP) / English (TE) into Norwegian
- recording date, recording place, iteration, recording environment, recording device, type of microphone, noise level, etc.

The labels of the recordings adhere to a structured format: **PROJECT_SPEAKER ID_LANGUAGE STATUS_TASK**, wherein:

- PROJECT corresponds to the project within which the data were collected (A for ADIM, C for CLIMAD)
- SPEAKER ID corresponds to a unique speaker ID consisting of 8 characters
- LANGUAGE STATUS represents the language in which the task was recorded and its status for the speaker (e.g., L1PL, L2EN, L3NO)
- TASK corresponds to the type of speech task recorded (e.g., TR, SR, WR, etc.). If a given task type was done more than once, numbers corresponding to their iterations have been added after TASK.

The LnNor corpus has been created to represent multilingual speech with a focus on L3/L*n* Norwegian learners as well as native controls of Norwegian, English and Polish. The corpus is designed to study linguistic variation in learners acquiring Norwegian as a foreign language in instructed and naturalistic settings. Additionally, a subcorpus of native speech patterns is provided to serve as a benchmark, against which the learners' productions could be compared. Furthermore, parts of the corpus contain word alignment with orthographic transcriptions of speech to facilitate subsequent analyses across various linguistic domains.

All speech samples were recorded with the use of Shure SM-35 unidirectional cardioid head-worn condenser microphones, using portable Marantz PMD620 solid state recorders with signal digitized at 48 kHz, 16-bit. This set-up was selected to minimize ambient noise and provide clear and focused recordings.

The **LnNOR corpus part 2** consists of 1739 annotated files from 153 speakers. The speakers included 113 L1 Polish, 18 L1 English, and 33 L1 Norwegian speakers. The total recording time is approximately 67 hours and the full size is 31 GB. The recordings in the released LnNor corpus part 2 cover data collected between 2023-2024.