

An Accurate Transcription Service for Belgian Dutch

Authors: Hugo Van hamme, KU Leuven, Belgium

Melina De Dijn, KU Leuven, Belgium

Abstract ID: 353

Event: Language Disorders 2026

Subject: 25. DELAD Special Session: Language Disorder Corpora

Presenter Name: Hugo Van hamme

Abstract

The FIIAT (Flemish Inclusive Infrastructure for Automatic Transcription) project establishes a state-of-the-art infrastructure for the automatic transcription of Belgian Dutch, directly addressing a critical bottleneck in research that relies on spoken documents such as interviews, focus group interactions, and daylong recordings. While existing international platforms and commercial tools perform reasonably well on standard speech, they frequently fail when encountering the colloquial or dialectal variation characteristic of authentic Flemish data, while also often struggling with child or elderly speech, and highly interactive exchanges. By providing a secure, locally run tool tailored to these diverse speaker groups, FIIAT ensures that transcriptions genuinely reflect natural language use, which is essential for the authenticity and validity of research and corpus creation for the humanities, social sciences, and biomedical fields.

Central to the project is a robust ethical and legal framework that addresses the significant risks inherent in modern AI-based language technologies. Speech data is identifiable and often sensitive. However, many researchers currently use online services without realizing that their data is processed on external servers where terms of service may permit the reuse their recordings. FIIAT eliminates these privacy and data security concerns by running the infrastructure locally at KU Leuven and at the Dutch Language Institute (INT), ensuring that data remains within a protected institutional environment.

The scientific impact of this infrastructure is broad and includes disciplines such as speech-language pathology, where capturing dysfluencies and non-standard speech without the implicit normalization that commercial tools often perform, is vital for language diagnostics. In the realms of developmental psychiatry and mental health, FIIAT facilitates the analysis of naturalistic verbal interactions and emotional dynamics through precise temporal alignment, allowing for integration with multimodal data like eye-tracking and physiological signals. Furthermore, it empowers researchers in linguistics and education to systematically document classroom interactions and regional dialect changes that were previously too time-consuming to analyze at scale.

Beyond its immediate utility, the infrastructure offers massive efficiency gains and fosters the creation of high-quality, FAIR-compliant (Findable, Accessible, Interoperable, Reusable) datasets. By automating the labor-intensive transcription process, FIIAT frees researchers to focus on core research tasks while providing advanced functionalities such as word-level confidence scores, word-level timings for fluency analysis, and customizable output formats.

Community building and dissemination are prioritized through the Leuven Interdisciplinary Language Institute (LILI), which acts as a central hub for ethical guidance and researcher training. LILI will also coordinate cross-disciplinary user needs to ensure the tool remains responsive to emerging requirements.

The service will go through several iterations as more training data is gathered. Its initial version will be active as of late 2026.