

OBJECTIVE

To identify reproducible **linguistic** and **acoustic speech biomarkers** of dementia by statistically analyzing structured clinical speech across multiple **DementiaBank** sub-corpora.

CONTRIBUTIONS

- Identifying **syntactic** and **acoustic** biomarkers (Parse Tree Depth, Verb Ratio) that independently replicate across two distinct clinical sub-corpora.
- Establishing a reliable, explainable, multimodal feature standard for automated speech-based cognitive screening, using a **statistical approach** with **FDR correction**.

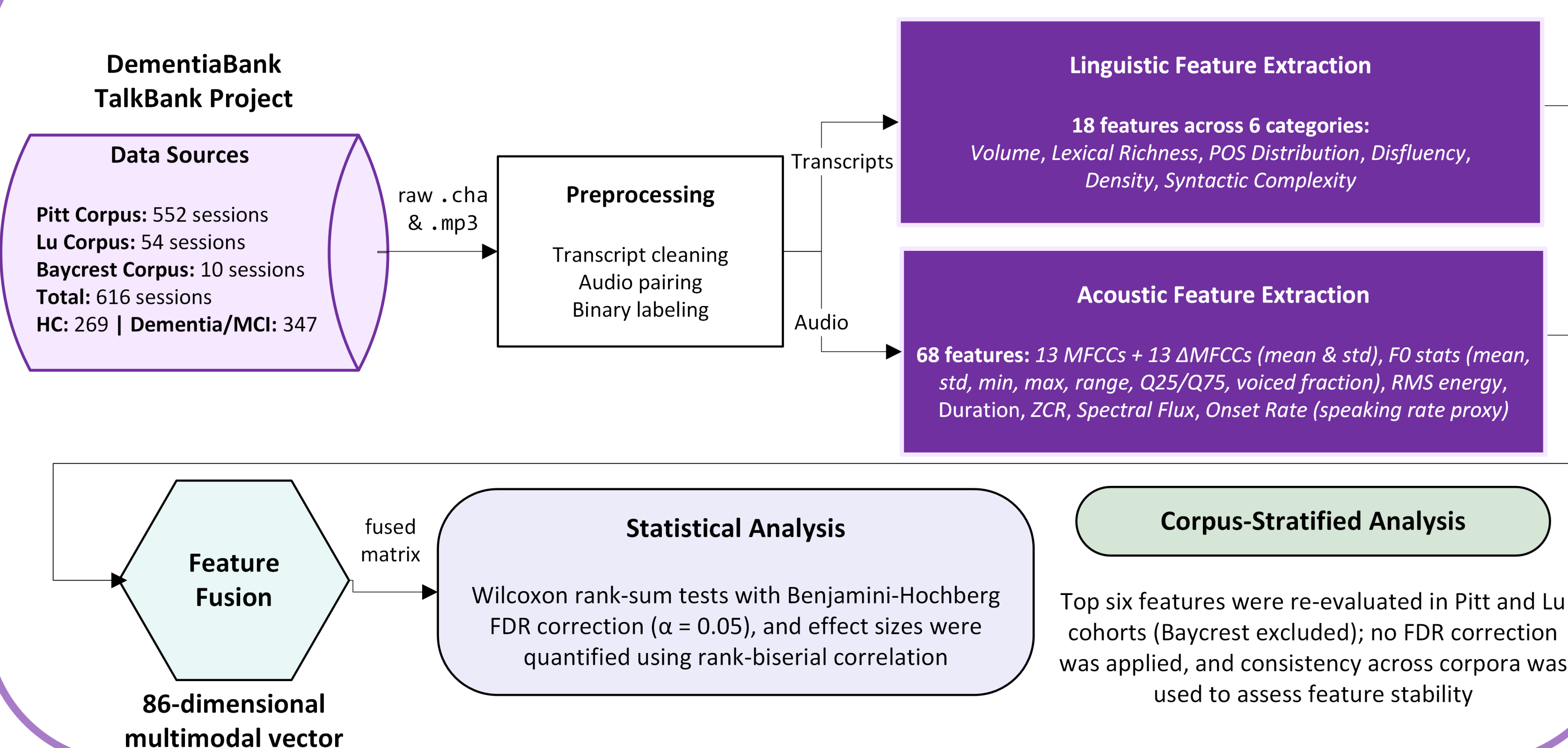
MATERIALS

- Dataset:** DementiaBank (TalkBank Project) — three sub-corpora: Pitt, Lu, Baycrest
- Total sessions:** 616 (Healthy Control: 269, Dementia/MCI: 347)
- Task:** Cookie Theft picture description (structured clinical elicitation)
- Diagnoses represented:** ProbableAD, MCI, PossibleAD, Control
- Format:** .cha (CHAT/CLAN transcripts) + paired .mp3 audio
- Toolbox used:** librosa (acoustic), spaCy en_core_web_sm (linguistic), scipy.stats (statistical testing), pandas/scikit-learn (fusion & imputation)

KEY TAKEAWAY

Parse tree depth and **verb ratio** are the strongest linguistic markers of dementia. Syntactic simplification and problems with verb retrieval can reliably tell apart AD/MCI from healthy controls. This finding comes from 616 real-world clinical speech sessions. It establishes a clear standard for automated cognitive screening.

METHODOLOGY



RESULTS

