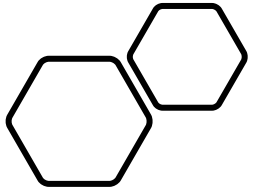


Rethinking Language
Computational approaches to speech and language assessment for clinical studies

How to Share Your Data in a GDPR-Compliant Way

4th CLARIN/DELAD workshop

Chiara Barattieri di San Pietro, PhD

University of Firenze, University of Brescia

About me

- Linguist with PhD in Psychology, Linguistic, and Neuroscience

- Now
  - Post-doc fellow @UNIFI
  - Lecturer @UNIBS

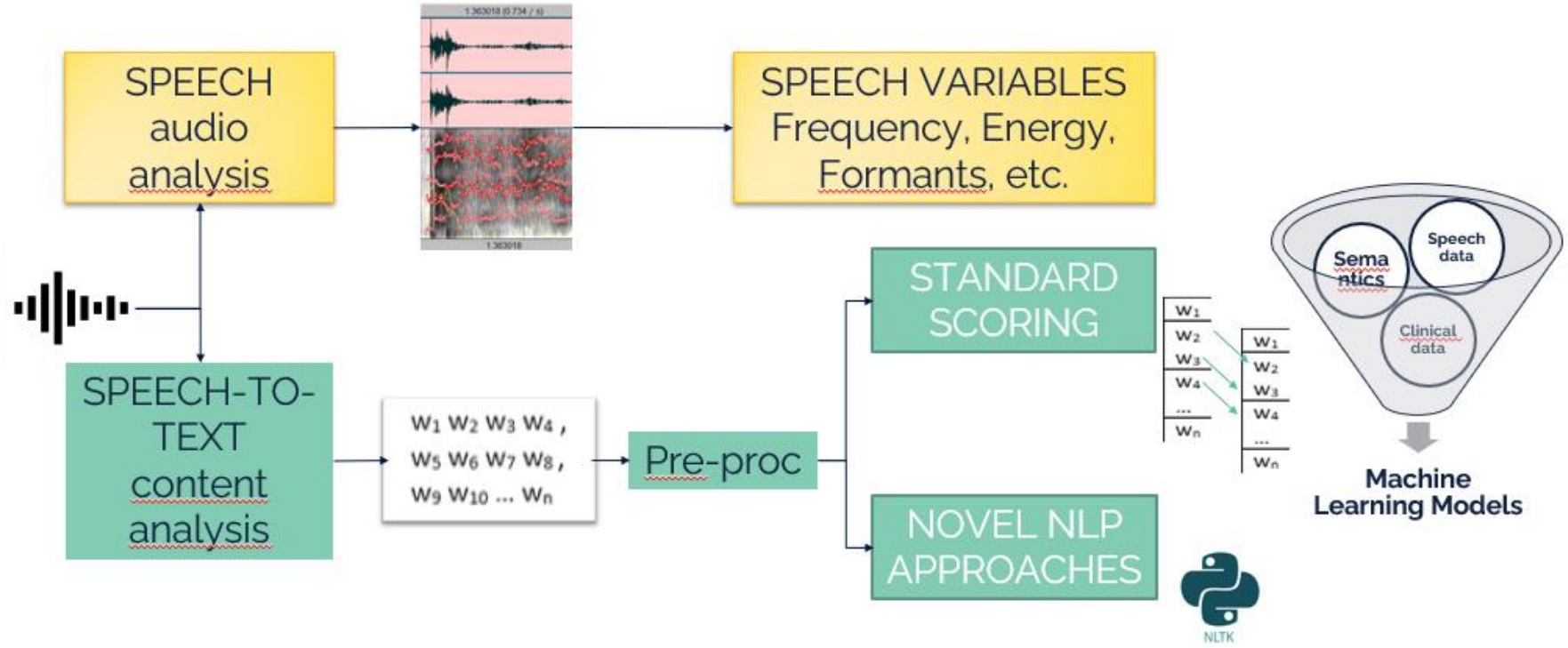
Why should we rethink speech and language assessment in the clinical practice?

- In a wide range of mental disorders, cognitive and affective symptoms are expressed in language and measured by assessing speech
  - Clinicians trained do that almost intuitively
  - There are standardized tests available, but
    - Paper-and-pencil
    - requires verbatim transcriptions and manual scoring
    - can only tap on surface phenomena

→ Time consuming and barrier to frequent/remote monitoring

Double path to language assessment

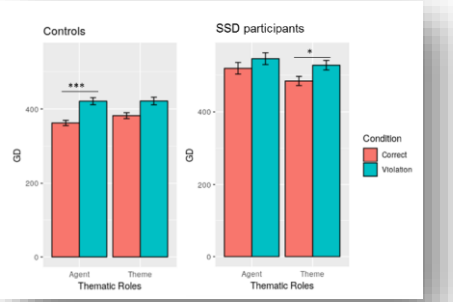
- Digital tools now available to model:
  - speech content: word vectors (LSA, word2vec, BERT, …)
  - Audio: spectral analysis to extract acoustic and paralinguistic aspects of voice



Our work so far
Analysis of syntactic and semantic content

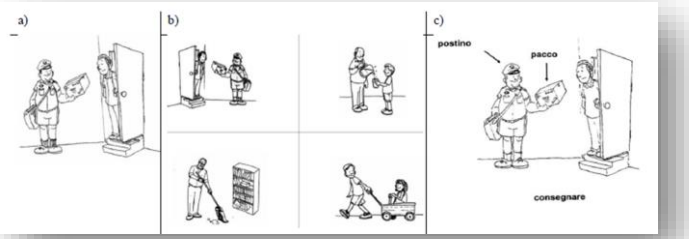
In schizophrenia some semantic violations are more tolerable than others - it depends on whether the grammatical subject is an agent or not!

Barattieri di San Pietro, C., de Girolamo, G., Luzzatti, C., & Marelli, M. (2022). Agency of Subjects and Eye Movements in Schizophrenia Spectrum Disorders. *Journal of Psycholinguistic Research*, 1-21.



Schizophrenia presents with language-specific deficits similar to those found in agrammatism

Barattieri di San Pietro, C. et al. (2022). Processing Argument Structure and Syntactic Complexity in People with Schizophrenia Spectrum Disorders. *Journal of Communication Disorders*, 106182



ML models based on DSM-derived measures of VF less accurate than models built on human-based ratings in classifying people with dementia.

Barattieri di San Pietro, C., Marelli, M., & Reverberi, C. (2021). Moving from Human Ratings to Word Vectors to Classify People with Focal Dementias: Are We There Yet?. CLiC-it.

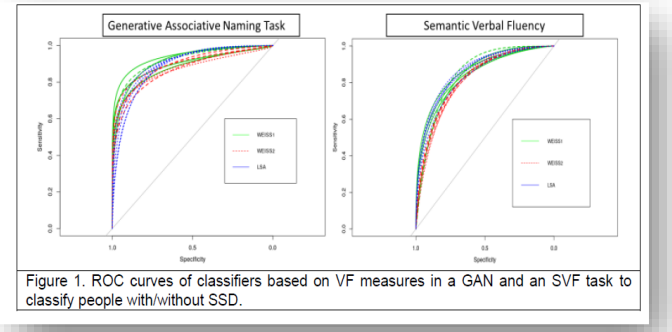


Figure 1. ROC curves of classifiers based on VF measures in a GAN and an SVF task to classify people with/without SSD.

What is going on
Clinical applications of voice and language analysis

- Application to very diverse clinical conditions
  - Psychiatric: BP, MDD, eating disorders…
  - Neurodevelopmental: ADHD, Autism…
  - Neurological: MCI, AD, strokes…
- With very diverse objectives:
  - Characterization of language in disorders
  - Classify subjects (patients vs controls, high risk vs low risk)
  - Predict prognosis (MCI to AD), conversion of high-risk subjects
  - Response to treatment
  - Monitor relapse and mood state (MDD and BP)
  - Remote monitoring
  - Personalized rehabilitation
- Analysing every possible audio feature
  - Must be clinically relevant!
  - Reduced variance of F0 in depression (monotonous speech)
  - Reduced speech rate (bradicynesia as negative symptom in BD, MDD, SZ)
  - Incoherence of responses (positive symptom in SZ)
  - Augmented speech rate (pressured speech – SZ and maniac state in BD)
  - Affective words (depressed mood in MDD and BP, negative symptom in SZ)

Bridging healthcare and IT - SPEAKapp

51 healthy participants, downloaded & installed app & completed all tests remotely.

- Data integrity: 99%
- Data quality: 90%
- WER: 10%
- Usability: 88.5%

→ Consistency of features: system variables in line with values in controlled environments

https://www.gatekeeper-project.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 857223

Ab.Acus

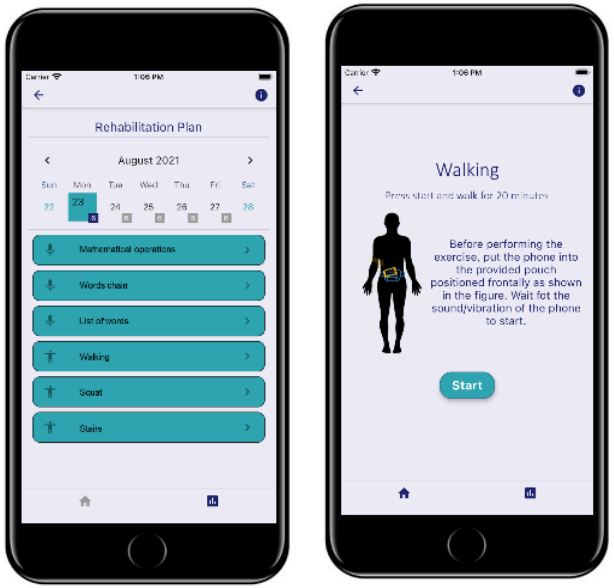
Clinical application
NLP, verbal performance and Digital Mental Health

- Objectives:
  - Primary: Identify language markers of symptoms severity in people with SMI (SZ, BD) via mobile app (SPEAKapp) to collect audio and standard diagnositc measures
  - Secondary: test acceptability of the system
- Methods
  - Observational, cross-validation study.
  - Test association of semantic and acustic features with patterns of symptom severity (for each clinical condition)
  - Participants: estimated 35+35
- Status
  - Pending EC approval. Data collection to begin in Jan ‘23 (approx)

FSE REACT-EU Competitive Research Grant Axis-IV DM 1062/2021: “Natural Language Processing in Digital Mental Health” - 2022-2024

PHY-CO! - Physical and Cognitive Rehabilitation Framework

- Cognitive and physical stimulation activities
  - Standard neuropsychological exercises
  - Physical exercises
- Status: ongoing
- Setting: rehabilitative outpatient clinics



This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 857188.

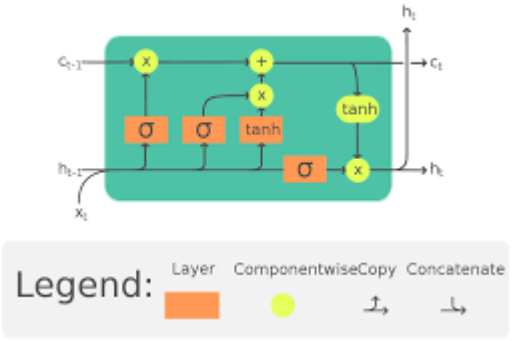
UNIVERSITÀ DEGLI STUDI FIRENZE

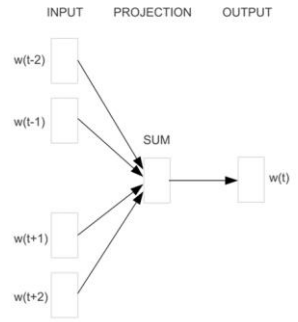
FORLILPSI
DIPARTIMENTO DI FORMAZIONE, LINGUE INTERCULTURA, LETTERATURE E PSICOLOGIA

Fine-tuning tools: Navigating textual ambiguity

MULTI WordNet

- Interplay between textual features and individual's ability to understand the text
- Computational measures of textual ambiguity
  - Lexical → n. synsets on MultiWordNet
  - Syntactic → surprisal index based on Long-Short-Term-Memory neural network
  - Pragmatic→ cosine distance btw metaphoric/literal expressions and their context, based on word2vec WEISS





Levelling up - Automated analysis of pragmatics

- Analysis of pragmatics in rehabilitation of neurological disorders (strokes)
- Data: N=30 patients (dx lesions), pre-post rehabilitation
  - MRI
  - EEG
  - neuropsychological evaluation & **audio samples**
- Status: data collection T1 completed, T2 ongoing

SAN CAMILLO IRCCS SRL

Going international
SWETALY mentoring programme

- Swedish-Italian cooperation on Health and Aging
- Secondary analysis on existing data from MCI patients
  - eye-movements
  - **audio recordings**
- Obj: find early digital markers of dementia

UNIVERSITY OF GOTHENBURG

SWETALY

My questions to the DELAD community

- Audio files
  - considered as health-related data («biosample») under GDPR and Italian law, and subjected to strict regulations
  - Some previously collected and more to come – how can **collaboration and replicability of results** be realized?
  - How can data be made available to the community, if at all possible? **Under which terms**?
- What about **anonymization tools**?
- Must be **paired with clinical data**

Thanks!

chiara.barattieridisanpietro@unifi.it

University of Florence
Prof. Christian Tarchi

Gothenburg University
Prof. Dimitris Kokkinakis

Ab.Acus srl
Maria Bulgheroni, Eng
Valentina Simonetti, Eng

IRCCS S. Lucia
Giorgio Arcara, MD
Sara Lago, MPsy

University of Milan-Bicocca
Prof. Marco Marelli
Prof. Claudio Luzzatti

IRCCS Fatebenefratelli
Giovanni de Girolamo, MD