
# Computational Methods Group

Pavia, Italia, April 8th 2024

Sunny Tang stang3@northwell.edu

Alban Voppel alban.voppel@mail.mcgill.ca

# Computational Methods Group

The mission is to:

develop,
harmonize,
validate computational methods

for understanding speech and language disturbances in psychosis.

26 members currently, regular meetings.

Mail either Sunny or Alban to be added to the mailing list

- stang3@northwell.edu
- alban.voppel@mail.mcgill.ca

# Current members:

| |
|---|
| Chiara Barattieri di San Pietro |
| Janna de Boer |
| Tuğçe Çabuk |
| Sylvia Ciampelli |
| Sunghye Cho |
| Yan Cong |
| Deanna Kelly |
| Federico Frau |
| Rui He |
| Wolfram Hinzen |
| Alexandra Korda |
| Esra Lenz |

| |
|---|
| Mark Liberman |
| Elena Lundaeva |
| Brian MacWhinney |
| Natália Mota |
| Caroline Nettekoven |
| Amir Nikzad |
| Matthew Nour |
| Lena Palaniyappan |
| Alberto Parola |
| Sameer Pradhan |
| Philip Resnik |
| Roberta Rocca |
| Sunny Tang |
| Alban Voppel |

# The past few meetings:

- Topics of interest inventoried and discussed
- Individual subgroups formed around specific topic areas and tasks.
- Identify own leadership and members, roles
- Registered on shared drive

# Group Processes:

- Individual subgroups will be formed around specific topic areas and tasks. These groups will identify their own leadership and assign roles, split up work needed to move toward the defined objectives.
  - Subgroups will be formalized/registered around an abstract describing the specific objectives.
  - Abstracts will be circulated across the larger working group and DISCOURSE for participation at different levels - contributing to analysis, data, etc.
- Function of the group at large: facilitate data sharing/access, coordinate with other working groups, forum for presentation and for discussion.
- Will liaise with other working groups within DISCOURSE:
  - Speechbank liaisons: Brian MacWhinney, Lena Palaniyappan, Mar Dominguez
  - Theory/Neurobiology liaisons: Wolfram Hinzen, Lena Palaniyappan

| Project nickname: | Proposal: | Interested Individuals + Affilitations: | (Prelim) Subgroup coordinator & contact information: | Comments/Status | Deliverable | Abstract |
|---|---|---|---|---|---|---|
| **Data Sharing Toolkit** | Create a data sharing kit with successful IRBs, outlines, etc. To be separate from Speechbank, as an opportunity to reach and teach other organizations. May be consistent with near future NIH opportunities | Sunghye Cho (UPenn), Sunny Tang, Brian MacWhinney, Mark Liberman, Phil Resnick | Sunny Tang (stang3@northwell.edu), Brian MacWhinney, Mark Liberman | | Commentary or other type of publication with publicly shared resources | |
| **Measure Stability** | Stability of measures across repeated measures and across multiple tasks | Chiara Barattieri di San Pietro (IUSS Pavia), Yan Cong (Purdue University), Silvia Ciampelli (UMCG), Alberto Parola (University of Copenhagen), Wolfram Hinzen, Sunny Tang (Northwell Health), Federico Frau (IUSS Pavia), Amir Nikzad (Northwell Health) | Wolfram Hinzen | Maybe be more than one group | Multiple publications | |
| **Course of Illness** | Stability vs. evolution of speech marker across the course of illness, chronic vs. acute vs. prodromal psychosis; Hypothesis - some trait and some state-related markers | Alban Voppel (McGill), Federico Frau, Silvia Ciampelli (UMCG), Sunny Tang, Alberto Parola (UCPH) | alban.voppel@mail.mcgill.ca | | | |
| **Semantic Space** | Cross-linguistic generalizability of shrinking semantic space via embeddings and graph analysis | Yan Cong (Purdue University), Tuğçe Çabuk (Bilkent University), Wolfram Hinzen (UPF), Sunny Tang | Wolfram Hinzen's group | | | |
| **Speech Graphs** | Cross-linguistic generalizability of lexical vs. semantically constructed speech graphs | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Tuğçe Çabuk (Bilkent University), Sunny Tang (Northwell Health), Amir Nikzad (Northwell Health), Natalia Mota, Federico Frau (IUSS Pavia) | Amir Nikzad (ANikzad@northwell.edu); Sylvia Ciampelli | | | |
| **Cross-linguistic acoustics** | Cross-linguistic acoustic analysis | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Alberto Parola (University of Copenhagen), Wolfram Hinzen (UPF), Sunghye Cho (UPenn), Federico Frau | Alberto Parola, Sunghye Cho | | | |
| **Topic modeling** | Topic modeling | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Sunny Tang, Federico Frau (IUSS Pavia) | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG) | | | |
| **Feature Space Reduction** | Evaluating approaches to feature space reduction | Sunny Tang (Northwell) | Sunny Tang (Northwell) | Deferred | | |
| **Multimodal models** | Multimodal models (text + audio) and cross-linguistic generalizability | Alberto Parola (University of Copenhagen), Federico Frau (IUSS Pavia) | Alberto Parola | Probably may be merged with project 7 | | |
| **Cross-Linguistic Toolkit** | Construct a table of languages and tools available | Sameer Pradhan (LDC) | Sameer Pradhan (LDC) | | Review paper? | |
| **Social Interactions** | Social determinants of language and interactions with language | Natalia Mota, Chiara BdSP (IUSS Pavia) | | Deferred | | |
| **Homomorphic encryption** | Computation with homomorphically encrypted data | Mark Liberman | Mark Liberman | Deferred | | |
| **Harmonized Pipeline** | Harmonized pipeline which could be applied across sites and across languages | Chiara BdSP (IUSS Pavia) | Chiara BdSP (IUSS Pavia) | Deferred | | |

DISCOURSE IN PSYCHOSIS

| Project nickname: | Proposal: | Interested Individuals + Affilitations: | (Prelim) Subgroup coordinator & contact information: | Comments/Status | Deliverable | Abstract |
|---|---|---|---|---|---|---|
| **Data Sharing Toolkit** | Create a data sharing kit with successful IRBs, outlines, etc. To be separate from Speechbank, as an opportunity to reach and teach other organizations. May be consistent with near future NIH opportunities | Sunghye Cho (UPenn), Sunny Tang, Brian MacWhinney, Mark Liberman, Phil Resnick | Sunny Tang (stang3@northwell.edu), Brian MacWhinney, Mark Liberman | | Commentary or other type of publication with publicly shared resources | |
| **Measure Stability** | Stability of measures across repeated measures and across multiple tasks | Chiara Barattieri di San Pietro (IUSS Pavia), Yan Cong (Purdue University), Silvia Ciampelli (UMCG), Alberto Parola (University of Copenhagen), Wolfram Hinzen, Sunny Tang (Northwell Health), Federico Frau (IUSS Pavia), Amir Nikzad (Northwell Health) | Wolfram Hinzen | Maybe be more than one group | Multiple publications | |
| **Course of Illness** | Stability vs. evolution of speech marker across the course of illness, chronic vs. acute vs. prodromal psychosis; Hypothesis - some trait and some state-related markers | Alban Voppel (McGill), Federico Frau, Silvia Ciampelli (UMCG), Sunny Tang, Alberto Parola (UCPH) | alban.voppel@mail.mcgill.ca | | | |
| **Semantic Space** | Cross-linguistic generalizability of shrinking semantic space via embeddings and graph analysis | Yan Cong (Purdue University), Tuğçe Çabuk (Bilkent University), Wolfram Hinzen (UPF), Sunny Tang | Wolfram Hinzen's group | | | |
| **Speech Graphs** | Cross-linguistic generalizability of lexical vs. semantically constructed speech graphs | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Tuğçe Çabuk (Bilkent University), Sunny Tang (Northwell Health), Amir Nikzad (Northwell Health), Natalia Mota, Federico Frau (IUSS Pavia) | Amir Nikzad (ANikzad@northwell.edu); Sylvia Ciampelli | | | |
| **Cross-linguistic acoustics** | Cross-linguistic acoustic analysis | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Alberto Parola (University of Copenhagen), Wolfram Hinzen (UPF), Sunghye Cho (UPenn), Federico Frau | Alberto Parola, Sunghye Cho | | | |
| **Topic modeling** | Topic modeling | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG), Sunny Tang, Federico Frau (IUSS Pavia) | Chiara BdSP (IUSS Pavia), Silvia Ciampelli (UMCG) | | | |
| **Feature Space Reduction** | Evaluating approaches to feature space reduction | Sunny Tang (Northwell) | Sunny Tang (Northwell) | Deferred | | |
| **Multimodal models** | Multimodal models (text + audio) and cross-linguistic generalizability | Alberto Parola (University of Copenhagen), Federico Frau (IUSS Pavia) | Alberto Parola | Probably may be merged with project 7 | | |
| **Cross-Linguistic Toolkit** | Construct a table of languages and tools available | Sameer Pradhan (LDC) | Sameer Pradhan (LDC) | | Review paper? | |
| **Social Interactions** | Social determinants of language and interactions with language | Natalia Mota, Chiara BdSP (IUSS Pavia) | | Deferred | | |
| **Homomorphic encryption** | Computation with homomorphically encrypted data | Mark Liberman | Mark Liberman | Deferred | | |
| **Harmonized Pipeline** | Harmonized pipeline which could be applied across sites and across languages | Chiara BdSP (IUSS Pavia) | Chiara BdSP (IUSS Pavia) | Deferred | | |

# Specific topic reports

1. Topic modelling in Psychosis
2. Cross-linguistic generalizability of lexical vs. semantically constructed speech graphs
3. Vocal and acoustic biomarkers

# Topic modeling in psychosis

Chiara Barattieri di San Pietro, Silvia Ciampelli, Federico Frau, Sunny Tang

# TM in psychosis

- Topic modeling (TM) help uncovering hidden thematic structures in text
- Mini literature review to:
  - describe the most frequently used topic modeling techniques and
  - their application in different psychiatric conditions.
- Searched for ([schizophrenia OR psychosis OR psychiatry] AND [topic modeling OR topic analysis]) in IEEE, Pubmed, EMBASE, Science Direct, and Springer Link.

# TM in psychosis  - Results

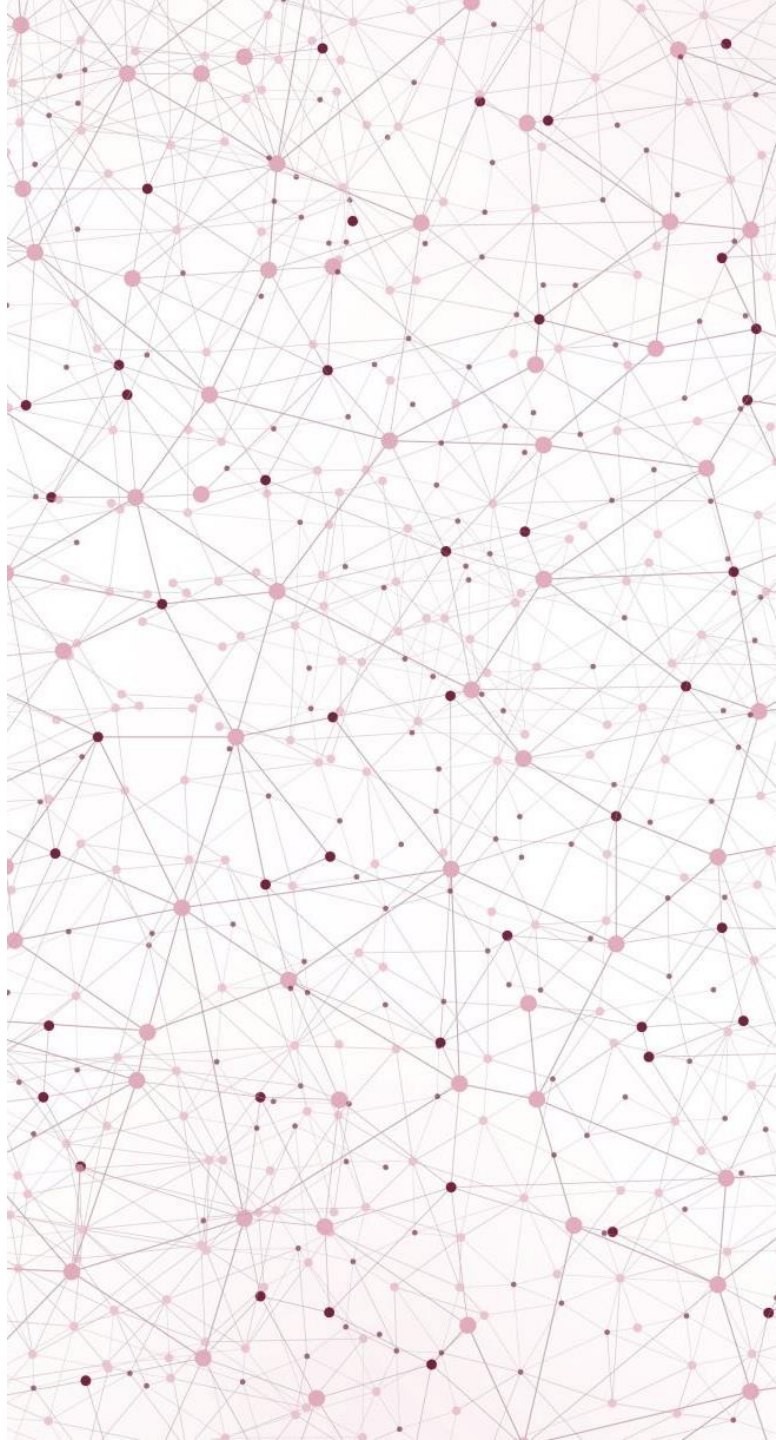
- Latent Dirichlet Allocation (LDA) is by far the most employed technique
  - followed by Non-Negative Matrix Factorization (NMF)
  - Others: Bidirectional encoding (BERTopic), NVDM-GSM, WTM-MMD, WTM-GMM, ETM, and BATM.
- Primarily applied to social media discussions, but also clinical records, and, although scarcely, psychotherapy sessions.
- Relevant topics identified:
  - food, living situation, lifestyle, symptoms, treatment experiences, energy level, people, and family.

# TM in psychosis – Future steps

- Challenges:
  - Substantial amounts of textual data to identify themes needed
  - Identified themes not always readily interpretable or clinically relevant
    - qualitative approach to complement any quantitative approach chosen (e.g., supervised learning, engage clinicians).
  - Variability related to the different data elicitation methods should be considered.
- Future research:
  - exploring temporal dynamics of topics, leveraging within-subject analyses to track symptom evolution over time.
  - Tested ability to predict outcomes like readmission rates and relapse, and distinguishing between FTD and non-FTD patients.

# Speech graphs and their relation to psychosis: a systematic evaluation across languages and graph types

**Amir Nikzad,* Silvia Ciampelli,* Chiara Barattieri di San Pietro, Federico Frau, Tuğçe Çabuk, Natalia Mota, Sunny Tang**



# Overview of graph types

| Level of linguistic analysis | Graph type | Nodes | Edges | Computational sub-processes | Key NLP Tool | Reference |
|---|---|---|---|---|---|---|
| Structural | Word-trajectory | Words or Lemmas | Sequential relation | None | Speech graph software | Mota et al., (2012) |
| Syntactic | Constituency parse tree | Words and Syntactic Categories | Syntactic relations | Sentence boundaries; Remove punctuation (for graph analysis); Keep punctuation (for parsing); Tokenization | Constituency parser | Ciampelli et al., (2023) |
| Semantic | Action/Predication | Arguments and/or Predicates | Differentiated semantic relations (action, predication, etc.) | Sentence boundaries; Semantic role labelling; Lemmatization | Semantic Role Parser | Nikzad et al., (2022) |
| Semantic | Semantic | Entities | Undifferentiated semantic relations | Sentence boundaries; Entity extraction; POS-tagging; Dependency parsing; Co-reference identification | Netts Python package | Nettekoven et al., (2023) |
| Pragmatic | Co-reference | Entities | Referential chain | NP identification; Recurrent and non-recurrent entities separation; Co-reference identification | Speech graph software | Palominos et al., (2023) |

# Analysis plan

1) map the availability of existing NLP tools to create different types of speech graphs;

2) standardize graph quantification techniques (eg., centrality, degree measures);

3) compare speech graph features with simpler NLP measures such measures of speech duration (e.g., word/sentence count) and lexical diversity (e.g., type-token ratio);

4) investigate feature redundancy/collinearity within and between graph types by statistical modelling;

5) compare the statistical strength of different graph types in relation to psychosis and its symptom dimension;

6) assess the cross linguistic generalizability of observed statistical relationships between speech graph features and psychosis.

# Vocal and Acoustic Biomarkers

Alberto Parola
(& Sunghye Cho)

# Cross-linguistic analysis of vocal and acoustic biomarkers

- This project explores the potential of machine learning (ML) in identifying vocal and acoustic markers for schizophrenia, with a particular focus on cross-linguistic generalization.
- The aim of the project is to explore how well acoustic patterns associated with schizophrenia generalize across language and context.
  - Test specific hypotheses about how language structures might affect vocal patterns and their interaction with speech tasks
  - Disentangle which features are more robust and generalize better across different languages
  - Test which methodological approaches are more robust and promising in terms of clinical applicability.
- Part of the project also involves investigating how the use of multimodal models – that is combining acoustic and textual features – could improve the generalization performance of the models.

| Project nickname: | Proposal: | (Prelim) Subgroup coordinators |
|---|---|---|
| **Data Sharing Toolkit** | Create a data sharing kit with successful IRBs, outlines, etc. To be separate from Speechbank, as an opportunity to reach and teach other organizations. May be consistent with near future NIH opportunities | Sunny Tang (stang3@northwell.edu), Brian MacWhinney, Mark Liberman |
| **Measure Stability** | Stability of measures across repeated measures and across multiple tasks | Wolfram Hinzen |
| **Course of Illness** | Stability vs. evolution of speech marker across the course of illness, chronic vs. acute vs. prodromal psychosis; Hypothesis - some trait and some state-related markers | alban.voppel@mail.mcgill.ca |
| **Semantic Space** | Cross-linguistic generalizability of shrinking semantic space via embeddings and graph analysis | Wolfram Hinzen's group |
| **Feature Space Reduction** | Evaluating approaches to feature space reduction | Sunny Tang (Northwell) |
| **Multimodal models** | Multimodal models (text + audio) and cross-linguistic generalizability | Alberto Parola |
| **Cross-Linguistic Toolkit** | Construct a table of languages and tools available | Sameer Pradhan (LDC) |
| **Social Interactions** | Social determinants of language and interactions with language | |
| **Homomorphic encryption** | Computation with homomorphically encrypted data | Mark Liberman |
| **Harmonized Pipeline** | Harmonized pipeline which could be applied across sites and across languages | Chiara BdSP (IUSS Pavia) |

Subgroups are making progress

Next group deliverables – abstracts in the form ☺

Date planner for the next meet
(also new agenda points!)

sign up for the mailing list

stang3@northwell.edu
alban.voppel@mail.mcgill.ca

# Questions and Input?