Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09768-2

EDITORIAL

Language resources for clinical linguistics: introduction to the special issue

Gloria Gagliardi[1] · **Marta Maffia**[2]

Accepted: 25 July 2024

Keywords Clinical Linguistics · Speech and language pathologies · Language resources

1 Clinical linguistics: past, present, and ongoing challenges in language sciences and linguistic resource collection

From the mid-1970s, a rich and growing body of research has delved into the linguistic characteristics associated with medical conditions. This highly interdisciplinary field of study, known as "Clinical Linguistics" since the pioneering studies of Crystal (1981, 1984, 1992), has gained increasing prominence and now stands as a fruitful and largely unexplored domain within the language sciences, poised for significant advancements at both theoretical and applied levels (Cummings, 2013).

Most children effortlessly acquire language during early neurodevelopmental stages, and, for most individuals, language competence remains stable throughout life. However, language acquisition is far from effortless for a significant number of kids, and a considerable number of adults experience language disturbances due to traumatic injuries, vascular accidents, or neurodegenerative diseases later in life (Damico et al., 2021). Therefore, as primarily suggested by Jakobson (8: p. 13), «for the linguist, who is concerned with the fully developed structure of language, its acquisition and dissolution cannot fail to provide much that is instructive».

✉ Gloria Gagliardi
gloria.gagliardi@unibo.it

Marta Maffia
mmaffia@unior.it

[1] Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

[2] Department of Literary, Linguistics and Comparative Studies, University of Naples L'Orientale, Naples, Italy

Published online: 30 August 2024

After more than 40 years, while the epistemological status of the discipline, as well as the common strategies for collecting and analyzing "pathological" speech and language, can be considered solved (Ball, 2021; Ball et al., 2008), many theoretical and methodological challenges remain open.

For instance: what contributions do corpus-based studies of speech and language disorders offer in terms of descriptive insights and theoretical frameworks, particularly in comparison to standardized testing? Can we establish "best practices" for the elicitation, collection, transcription, and annotation of clinical linguistic datasets? Not least in importance, in the era of big data, clinical linguistics faces persistent challenges with obtaining reliable data and coordinating efforts in resource collection, largely due to ethical committee approvals that restrict the shareability of clinically valid corpora. Consequently, research groups can only develop limited linguistic resources tailored for specific purposes authorized by research ethics boards. Thus, since patients' recordings, transcripts, and written productions are "special categories of personal data" subjected to stringent data-protection safeguards, is data reusability possible? And what fresh challenges do the new technologies, especially Natural Language Processing (NPL) techniques and Artificial Intelligence, raise?

In this complex framework, this special issue aims to focus on the methodologies for collecting linguistic resources – corpora, lexica, database, ontologies – and developing (computational) tools for the annotation of atypical language due to neurodevelopmental or acquired disabilities.

The following topics of interest were addressed by the call for papers:

- Corpus-based research in speech-language pathology: acquisition, creation, annotation.
- Identification and use of language resources for clinical applications (e.g., screening, diagnosis, monitoring, and phenotyping of language-related disorders).
- Design of annotation tools for the analysis of atypical verbal productions.
- Legal and ethical aspects of language technology: good practices for the collection, treatment, storage, sharing, and dissemination of linguistic data from healthcare settings.

We began developing this proposal in mid-2021. In addition to issuing an open call for papers, we reached out to renowned researchers in the field. Ultimately, eight papers were selected for inclusion, covering a diverse array of topics: from the proposal of new corpora for the observation of language production in certain disorders (e.g. dementia and schizophrenia), to new analytical approaches and tools for assessing language (pragmatic) skills, to methods for harnessing NLP techniques to support clinicians, even for the early diagnosis or detection of symptoms of various pathological conditions, such as breast cancer or depression. A brief overview is provided in Sect. 2.

2 The contents of this issue

Undoubtedly representing a good practice in response to the question posed above about the shareability and re-usability of speech and language data, the opening contribution by Lee and colleagues offers a presentation of the DELAD initiative. Launched in 2015 by Professors Martin Ball and Nicole Müller, it primarily aims at providing linguists and clinicians with an international platform for archiving and sharing speech data, thereby improving research on communication disorders and treatment practices. The document also provides information and recommendations on three sensitive issues in the field of clinical phonetics and linguistics: the annotation procedure, describing the most used tools and techniques and suggesting the use of uncertainty markers and continuous dimensions instead of categorical labels; participants' consent and data storage and access, for which precise ethical guidelines and useful examples are proposed; and data protection risk assessment, reporting an educational role-paly activity that can also be used with students. Links to all materials on the DELAD website are included in the text.

The following two papers present new resources for the analysis of disordered speech productions.

In the article by Gkoumas et al., a fine-grained longitudinal and multi-modal corpus of speech of individuals with various forms of dementia and healthy speakers is introduced. The corpus contains spoken interactions and handwritten or typewritten productions collected in a natural non-clinical setting and elicited through reminiscence materials from the two cohorts of participants. Concerning methodology, a tablet application was used for task administration and data collection, which also allowed for the acquisition of extra-linguistic information such as pen strokes and keystrokes. Preliminary results from a series of experiments are reported, showing higher longitudinal language variations in people with dementia compared to the control group.

The paper by Raso and colleagues focuses on speech in schizophrenia, presenting a Brazilian spoken corpus collected as part of the C-ORAL-ESQ (Corpus Oral de Esquizofrênicos) project at the Federal University of Minas Gerais. The main objective of the project is to observe the effects of cognitive impairment on the management of information structure and speech prosody. The clinical and ethical criteria applied for the recruitment of participants and the data collection procedure are described in detail, as well as the corpus features and the transcription and annotation methods. The results of some pilot studies are also presented, which reveal a simplification of schizophrenic speech both cognitively and prosodically, compared to non-pathological speech.

Cangemi et al. also report a study on schizophrenic speech, applying a new analytical method to a pre-existing corpus. The authors explore the possibility of observing content-free speech activity records with the aim of obtaining interactional specific profiles of individuals with schizophrenia, while adhering to constraints on ease of annotation and privacy requirements. Through visual inspection of the distribution of silences and vocal activity, the authors applied several types of duration-based analyses on speech samples from multiple interviews between therapists and three patients in the CIPPS corpus. The results show high inter-subject variability, attributed to

differences in patients' behavior and symptoms on the schizophrenia spectrum, and low intra-subject variability, allowing for the definition of distinctive communication styles associated with this disorder.

In the contribution by Bischetti and colleagues, a new method for the rapid telematic assessment of receptive and productive pragmatic skills in the Italian-speaking population is proposed. Described as potentially applicable to a wide range of clinical conditions and communicative impairments, the new APACS (Assessment of Pragmatic Abilities and Cognitive Substrates) Brief Remote test is modelled on the already validated in-person APACS test, simplified and adapted for the online administration procedure. Grounded in Gricean theory, the test includes 18 original items assessing discourse and non-literal language understanding. The test was administered to a sample of 141 healthy participants, demostrating adequate results in terms of reliability and validity measures. Additionally, an alternative form of the test is presented to allow for monitoring and follow-up.

With the ultimate goal of supporting the early detection of cancer through NLP, the study by Hepsağ et al. consists in the application of pre-trained language models (BERT and DistilBERT) and basic machine learning algorithms on a novel Turkish dataset of 62 mammography reports, previously labeled as malignant or benign by an expert in the diagnosis of breast cancer. The text classification performances of the models were compared and evaluated on the original dataset and on augmented and balanced versions. Moreover, an ensemble classifier was used to improve the results of the pre-trained BERT models. The results are extensively detailed, demonstrating that BERT and Hard Voting outperform other machine learning models in breast cancer diagnosis from Turkish radiology reports.

Staying within the NLP domain, the work of Farruque and colleagues describes a Semi-Supervised Learning (SSL) framework designed to detect cues of depressive symptoms from the Twitter timelines of self-declared users. The SSL framework is constituted by several iterative steps that begin with training a Depressive Symptoms Detection (DSD) language model on clinician-annotated data and proceed with various phases of data collection, evaluations and fine-tuning. The main goal of the research is to create the largest self-curated Depressive Tweets Repository (DTR). Each step of the data harvesting and model training and retraining processes are presented and accurately discussed in the paper.

Finally, turning to the field of robotic-assisted surgery, Bombieri et al. offer a description of the first annotated resource aimed at enhancing surgical natural language understanding. A dataset of sentences from textbooks and academic papers on different robotic-surgical procedures was manually annotated using an adapted version of the PropBank semantic labels, enriched to include domain-specific language. The new Robotic-Surgery Procedural Framebank (RSPF) and the methodology used for multilevel annotation of the corpus are outlined. Moreover, the RSPF and the annotated dataset are publicly released in this contribution, to promote further research in this direction and to enable the benchmarking of Semantic Role Labeling tools in the robotic-surgery domain, which is still poorly explored for natural language understanding.

3 Concluding remarks

In closing, we trust that this concise overview of each article will assist readers in exploring this special issue and encourage them to delve into the individual contributions.

We would like to take this opportunity to thank all the authors who contributed with their works, and all the reviewers who generously provided their time and expertise to help us select and improve the best submissions. We extend our heartfelt gratitude to Nicoletta Calzolari, Nancy Ide, and Sara Goggi for their invaluable guidance throughout the entire editorial process.

Author contributions Both authors have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Funding Neither author received financial support for writing this introduction.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ball, M. J. (Ed.). (2021). *Manual of clinical phonetics*. Routledge.

Ball, M. J., Perkins, M. R., Müller, N., & Howard, S. (Eds.). (2008). *The Handbook of Clinical Linguistics*. Blackwell.

Crystal, D. (1981). *Clinical Linguistics*. Springer.

Crystal, D. (1984). *Linguistic encounters with Language Handicap*. Blackwell.

Crystal, D. (1992). *Profiling linguistic disability*. Edward Arnold.

Cummings, L. (2013). Clinical linguistics: A primer. *International Journal of Language Studies*, *7*(2), 1–30.

Damico, J. S., Müller, N., & Ball, M. J. (2021). *The Handbook of Language and Speech Disorders (Second Edition)*. Hoboken (NJ): Wiley-Blackwell.

Jakobson, R. (1941). *Kindersprache, Aphasie und allgemeine Lautgesetz*. Uppsala: Almqvist & Wiksell. (English translation, 1968, *Child language, Aphasia, and Phonological Universal*. The Hague/Paris/New York: Mouton Publishers).