Disordered Speech Resources for Polish
Potential and Perspectives

Anita Lorenc [1]
Katarzyna Klessa[2]
Ewa Wolańska[1]
[1]Institute of Applied Polish Studies, University of Warsaw, [2]Institute of Linguistics, Adam Mickiewicz University in Poznań

DELAD Initiative Workshop, 15-17.11,2017, Cork, Ireland

Our Team

Warsaw – Poznań

& more

Anita Lorenc - clinical phonetician, speech, language & hearing therapist

(Institute of Applied Polish Studies, University of Warsaw)

Katarzyna Klessa - phonetician involved in speech technology projects & speech resources/tools development, corpus annotation and other.

(Institute of Linguistics, Adam Mickiewicz University in Poznan)

Ewa Wolańska - linguistic scientist, speech and language therapist, clinical speech-language pathologist

(Institute of Applied Polish Studies, University of Warsaw)

Disordered speech & reference datasets

comparisons between languages but in the first step within language

Multi-channel recordings of controlled speech

- As a result of an on-going project managed by Anita Lorenc, a dataset of multi-channel audio recordings has been collected (Project: ***Polish Language Pronunciation. Analysis Using 3-dimensional Articulography*** financed by The Polish National Science Centre, ID: 2012/05/E/HS2/03770)
- Recordings obtained with a 16-channel microphone array with a dedicated audio recorder, additionally using electromagnetic articulograph AG 500, and a vision component consisting of 3 high-speed cameras.
- Recording scenarios were aimed to enable the empirical verification of **Polish contemporary pronunciation.**
- The dataset contains only utterances produced by healthy speakers without any speech disorders (verified by experts) thus it can be used as **a reference corpus** in research on disordered speech.

Other “reference” corpora

As far as databases of speech of healthy speakers we also have access to other resources includings data for varying speaking styles, e.g.:

- Paralingua corpus of task-oriented dialogues & affective speech
- an on-going work on a corpus of conversational speech of teenagers from the Polish-German borderland area (http://borderland.amu.edu.pl)
- language archives for endangered languages and dissemination of knowledge about endangered or under-resourced languages (inne-jezyki.amu.edu.pl, languagesindanger.eu)
- & more.

Child speech – vocabulary lists

Available datasets collected earlier by Anita Lorenc include also speech recordings (vocabulary list) produced by:

- 20 children with a **deep hearing disability** aged between 8 and 12
- 10 children with a **correct hearing** as a control group.

Patients with pragnosia, aphasia and dysarthria

On-going work: **a corpus of speech data** collected for a PhD thesis under supervision of Anita Lorenc:

- data on speech of patients with **pragnosia, aphasia** and **dysarthria**;
- patients recorded on location in hospitals (non-studio conditions);
- conversational speech & more.

Research on apahasia, neuroliguistics & more

- Ewa Wolańska, as the Head of the Postgraduate Speech and Language Therapy Study at Warsaw University conducts research focusing on issues related to **aphasia, neurolinguistic**s and speech and language disorders in **neurodegenerative diseases**, particularly in **Alzheimer's** disease.
- Ewa also specializes in **neurological diagnosis** and therapy of adults with **CNS damage**, **stroke**, traffic **accidents**, **cranio-cerebral traumas**, brain **tumors** and other neurological diseases.
- No datasetgs ready but a possibility to design & develop ones.

Possible areas of interest

- neurogenic disorders driven by different factors (e.g., dysarthria, apraxia of speech, aphasia, pragnosia),
- neurodegenerative diseases (e.g., logopenic variant of Alzheimer's disease – LvAD, apraxic variant of Alzheimer's disease – AvAD, non-fluent variant of primary progressive aphasia – nfvPPA),
- developmental child speech disorders,
- hearing impaired child speech disorders.

Possible areas of interest

vs.

Areas of interest by DELAD

?

- neurogenic disorders driven by different factors (e.g., dysarthria, apraxia of speech, aphasia, pragnosia),
- neurodegenerative diseases (e.g., logopenic variant of Alzheimer's disease – LvAD, apraxic variant of Alzheimer's disease – AvAD, non-fluent variant of primary progressive aphasia – nfvPPA),
- developmental child speech disorders,
- hearing impaired child speech disorders.

Possible locations for data collection

hospital environment & laboratory conditions

Possible locations (1)

As a part of internship conducted by **University of Warsaw and Warsaw Medical University** in the course “General and Clinical Speech-Language Pathology” we have access to Warsaw's clinical hospitals, rehabilitation clinics and selected speech therapy clinics on cases of:

- neurogenic disorders driven by different factors,
- neurodegenerative diseases,
- developmental child speech disorders,
- hearing impaired child speech disorders.

Possible locations (2)

- Speech Therapy Guidance Service, Speech Therapy Center Warsaw University.
- Phonetic Laboratory, Speech Therapy Center, Warsaw University.

DELAD experiences / suggestions regarding the recording locations

Some tools & techniques used

- Existing software(e.g. Praat, ELAN)
- Existing therapy practice
- New tools and solutions both in the domain of data processing and diagnosis

Speech annotation - http://annotationpro.org/

- We are involved in the development of **Annotation Pro**, a tool for multilayer annotation of both linguistic and paralinguistic features of speech.
- It is possible to use the tool in combination with other popular tools such as Praat, Transcriber, ELAN or Wavesurfer thanks to its **import/export** functions.
- Annotation Pro has so far been used for multilayer annotation of several speech corpora, as well as **annotation mining** for data imported from external corpora.
- The work included among other analysis of temporal variability in utterances of French **healthy and dysarthric** speakers (for TYPALOC corpus, Bigi et al., 2015).

Database systems & dissemination of knowledge

Our experiences include creation of relation databases for the purpose of corpus annotation management, e.g. for the needs of:

- the “reference” corpora mentioned above (Paralingua corpus, speech of teenagers from the Polish-German borderland area, language archives for endangered languages);
- speech technology corpora (automatic speech recognition corpora, lexical databases)

Possible types of data

audio

- **audio data** in terms of all examined people – they can be relatively easily collected in clinical conditions as well as (for the majority of speakers) in the laboratory ones

Possible types of data

video

- **video recordings** - studio quality; possible on condition that a patient gives his/her consent also to video image recording, and that their medical condition is stable enough to enable participation in the recording at the Centre of Speech-Language Pathology at the University of Warsaw.

Possible types of data

articulographic

- selected patients can be examined with the use of **articulograph** (AG501 model, by Carstens Company) – only selected persons due to the fact that the examination is time-consuming and costly in terms of both obtaining and analysis of samples.

Questions (1)

Since we are new to the DELAD initiative our questions about guidelines include but are not limited :) to the following:

- tools to develop / adjust?, pronunciation questionnaires for the purpose of pronunciation studies? word-lists? spontaneous / controlled speech?
- recording procedures & scenarios
- the choice of speakers, criteria, speakers' age (adults / children?)
- kind of disorders (realisation of phonological subsystem, lexical-semantic, syntactical…), and/or communicative skills, the level of dysfunctions?

Questions (2)

- the size of speech samples per speaker,
- technical standards of audio and video recordings,
- the principles of segmentation and annotation of speech signal etc.,
- exchange of experience with scientists having conducted the required kind of research before,
- access to example analysis of data received from the patients,
- any available tools? automatic transcription or other?
- transcription of disorders & its level of detail?

Summary

Potential: existing datasets of disordered speech and standard Polish, tools for corpus annotation & annotation mining, expertise in speech therapy & analysis, contacts with both therapists and researchers.
Perspectives: possibility of extensions or new developments after clarifying the doubts / questions, possible support from CLARIN.PL (Katarzyna Klessa has recently become a member of the scientific council of CLARIN.PL).

Thank you!

Contact e-mails:

anita.lorenc@uw.edu.pl

klessa@amu.edu.pl

Thank you!

Acknowledgements:

- Polish Language Pronunciation. Analysis Using 3-dimensional Articulography financed by The Polish National Science Centre, ID: 2012/05/E/HS2/03770
- INNET (Innovative Networking in Infrastructure for Endangered Languages) European FP7 Project on endangered languages (languagesinmdanger.eu website and resources).
- Poland's linguistic heritage archive, grant No. 11H 11 001480 grant from the Polish Ministry of Science and Higher Education within The National Programme for the Development of Humanities in the years 2012-2014 (inne-jezyki.amu.edu.pl)
- Paralingua corpus. Grant no. O R00 0170 12 supported from the financial resources for science in the years 2010–2012 as a development project.
- Language of Boundaries and Boundaries of Language. Grant from the Polish Ministry of Science and Higher Education within "The National Programme for the Development of Humanities" in the years 2014-2016 (Polish-German multimodal corpus, borderland.amu.edu.pl).
- Project no. UDA-POIG.08.01.00-24-228/09-00 funded by the EU. *Portal edukacyjny polskiej fonetyki stosowanej w zakresie normy i patologii mowy,* Program Operacyjny Innowacyjna Gospodarka.