DEVELOPING A SPEECH CORPUS FOR A LESS-RESOURCED ROMANCE LANGUAGE. THE CASE OF CONTEMPORARY STANDARD ROMANIAN

Objectives. The aim of this presentation is to describe the recent development of a Romanian speech corpus suitable for phonetic analysis. We will be covering the tasks related to data collection (experimental setup) and data processing (annotating and transcription).

Introduction. The redirecting of speech technologies such as automatic alignment, speech synthesis, data mining, natural language processing into the Humanities has brought a reciprocal gain, expanding interdisciplinary research, linking text, speech and language technologies. Digital Humanities are bringing forth new methods and tools into the study of language variation and sound change [1]. Working on large data sets of annotated speech has proven to be of high relevance especially in phonetic research [2] and linguistics in general by leading to a better understanding of the evolution within a language [1]. As a result, modern phonetic research has drawn close to speech technologies [2]. Among European languages, Romanian is described as a less-resourced language, not benefiting from a well-established representation in domains such as speech processing, or large data scale analysis [3]. Easy to access speech corpora prove hard to find and explore [4], particularly from a phonetic and phonological perspective. Unlike the existing oral corpora for the main Romance languages [5, 6, 7, 8], researchers interested in Romanian casual speech data have a rather limited number of options. Most of the existing national oral annotated corpora, such as CORV [9], IVLRA [10], ROVA [11], only allow access to the transcribed material (adapted for pragmatic studies), without the corresponding audio files, thus limiting the analyses conducted at the interface between phonetics and phonology. The largest collection of dialectal recordings, gathered within the *The Phonogramic Archive of the Romanian Language* [12], remains an in-house resource in terms of open access to the digital files resulting from the reel-to-reel tapes digitization. The recorded material is depicted in dialectal monographs, anthologies, glossaries and dictionaries. Recent corpora, such as CoRoLa [13], mainly cover written text data acquisition and processing, while exploring the existing audio files is limited to the option of searching for only a certain word or lemma. The user is unable to access the entire recordings or the background information (speaker age, gender, social status). As a result, in this presentation we address the pending need to develop tools and linguistic resources for less represented Romance languages.

Corpus description. The scope of our research project is to offer a high-quality open access monolingual speech corpus of contemporary standard Romanian, accounting both for read and casual speech. This corpus will allow for an in-depth examination of language variation ranging from phonetic, prosody, semantics and morphology, to syntax and pragmatics. The data thus acquired can contribute to general comparative Romance linguistics, a domain where the Romanian language is often misrepresented. Since the recording and transcription of the corpus are one of the key objectives of our current postdoctoral research project aimed towards developing linguistic resources for an under-resourced Romance language, there are certain restrictions in terms of time (2 years) and human resources (limited to the postdoctoral researcher). As a result, we focus our analysis on 12 speakers (6 female, 6 male), ages 30 to 45, sharing the same educational and geographical background (representative of the southern dialect on which the standard language is based on). The corpus is recorded in a phonetic laboratory, within a sound-attenuated room of approximately 2x3m, equipped with a studio quality microphone connected to a laptop through an external audio interface (44100 Hz, Mono). Speaker-metadata is provided for each recording. All participants sign an agreement following the GDPR norm. It is also important to mention the fact that speakers are not wearing surgical masks or respirators. In terms of controlled speech, participants read randomized stimuli (the target word is placed in a carrier sentence “Zic__tare.”; “I say__loud.”), at a normal speech rate, going through the entire set three times. In order to control for speech rate and final intonation contours, each utterance appears on a screen for 3.5s, while the duration between slides is limited to 2s. This experiment focuses on extracting relevant data pertaining to VOT measurements of Romanian voiceless /p, t, k/ and voiced stops /b, d, g/ (placed in monosyllabic words), the 7 cardinal vowels /i, ɨ, u, e, ə, o, a/ in stop – vowel – bilabial /p/ logatomes (thus extending the analysis conducted in [14]), as well as fricatives /s, z, ʃ, ʒ, f, v, h/ in initial, medial, and final positions (in logatomes following the structure fricative – vowels /i, a, u/ – voiceless bilabial /p/, intervocalic position, and /p/ – vowel – fricative, respectively). Information regarding the friction noise is also extracted from the 3 affricates /ʦ, ʧ, ʤ/ present in the language following the same contexts. This selection for the controlled experiment is motivated by the absence of acoustic data for standard Romanian pertaining to the aforementioned topics. In terms of spontaneous speech, participants are required to undertake a monologue task (it does not involve speaker overlap, it allows for full data recovery, and is also faster to transcribe and annotate). At the beginning of the experiment, each candidate received a handout containing the three main conversational topics, namely *pursuits* (refers to present activities) – in my spare time, travelling, what I like, what I dislike, *memories* (refers to past activities) – from childhood, life lessons, and *forthcoming projects* (personal and professional). Fifteen minutes are given prior to the recording session during which participants can write down their ideas in relation to the topics proposed using only key-words. Participants can move freely from one topic to another, no order is pre-imposed, the goal being to emulate natural, (semi)spontaneous speech and, at the same time, allow for future cross-examination based on comparable data collection. The corpus is undergoing manual transcriptions. All transcriptions are carried out in Praat. The first tier contains the orthographical transcription paired with a board phonological transcription present on the second tier. In terms of annotating the speech data, the system accounts for: silent pauses ("#"), pause fillers ("@" placed in front of the filler, we distinguish between vowel @ă, nasal @m, vowel + nasal coda @ăm), hesitations ("%") and repetitions ("+"), truncation ("-", alongside the orthographic form restoration so as to facilitate the search of specific lexical items), elision (the deleted segment or syllable is marked between round brackets – the most frequent phenomena is that of the definite article *-l* deletion; our findings can further extend the research on Romanian speech variation conducted in [15]), laughter, lengthening of segments. Indistinguishable speech is marked between square brackets, while code switching appears between curly brackets. The label "^" was used for respiration noises which are delimited in the TextGrid. Names are written in capital letters and mispronunciations (including colloquial number pronunciation) are marked with "*".

Conclusions. In this presentation, we offered an insight on a recently developed open-access Romanian speech corpus which is set to be released in the second part of 2022. The corpus is designed in order to facilitate research at the interface between phonetics and phonology by looking at inter- and intra- speaker variability with respect to various reduction processes found in connected speech. This analysis opens up numerous discussions with respect to modelling gradient phonetic and phonological phenomena, leading to an in-depth study of linguistic variation and sound change. The material recorded also addresses a wide variety of scientific issues relevant for both linguistic and ASR applied research. The corpus can offer a new insight on a Romance language less studied on connected speech data, but with an increasing demand of digital learning tools.

References. **[1]** Ohala, J.J, 1996, "The connection between sound change and connected speech processes", *Arbeitsberichte* (AIPUK 31) Universität Kiel, pp. 201-206. **[2]** Adda-Decker, M., 2006, "De la reconnaissance automatique de la parole a l'analyse linguistique des corpus oraux", Paper presented at *Journées d'Étude sur la Parole* (JEP 2006), Dinard, France, 12–16 June. **[3]** Trandabat, D., E. Irimia, V. Mititelu, D. Cristea and D. Tufis, 2012, *The Romanian Language in the Digital Age*, META-NET White Paper Studies, Springer. **[4]** Mîrzea-Vasile, C., 2017, "Corpusurile de limba română și importanța lor în realizarea de materiale didactice pentru limba română ca limbă străină", *Romanian Studies Today*, I, București, Editura Universității din București, pp. 74-95. **[5]** Cresti, E., F. Bacelar, A.M. Sandoval, J. Veronis, P. Martin and K. Choukri, 2004. " The CORAL-ROM CORPUS. A multilingual resource of spontaneous speech for Romance languages". Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, pp. 575-578. **[6]** Torreira, F., M. Adda-Decker and M. Ernestus, 2010, "The Nijmegen corpus of casual French", *Speech Communication*, 52(3), pp. 201-212. **[7]** Torreira, F. and M. Ernestus. "The Nijmegen corpus of casual Spanish." *Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010. **[8]** Mereu, D. and A. Vietti, 2021, "Dialogic ItAlian (DIA): the creation of a corpus of Italian spontaneous speech", Speech Communication 130:1-14. **[9]** Dascălu Jinga, L., 2002, *Corpus de română vorbită (CORV). Eșantioane*, București, Oscar Print. **[10]** Ionescu-Ruxăndoiu, L. (coord.), 2002, *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, București, Editura Universității din București. **[11]** Dascălu Jinga, L. (coord.), 2011, *Româna vorbită actuală (ROVA). Corpus și studii*, Academia Română, Institutul de Lingvistică "Iorgu Iordan – Al. Rosetti". **[12]** Șuteu, V., 1958, "Arhiva fonogramică a limbii române", *Fonetică și dialectologie*, I, pp. 211-219. **[13]** Barbu Mititelu, V., D. Tufiș and E. Irimia, 2018, "The Reference Corpus of the Contemporary Romanian Language (CoRoLa)", *Proceedings of LREC 2018*, Japan, pp. 1178-1185. **[14]** Renwick, M. EL., 2014, *The Phonetics and Phonology of Contrast*. De Gruyter Mouton. **[15]** Vasilescu, I., I. Chitoran, B. Vieru, M. Adda-Decker, M. Candea, L. Lamel and O. Niculescu, 2019, "Studying variation in Romanian: deletion of the definite article -l in continuous speech", *Linguistics Vanguard – de Gruyter*, 5(1), pp. 1-12.