SCIENTIFIC REPORT
on the project
Acquiring and exploring an oral contemporary spoken Romanian corpus for linguistic purpose
(Project code: PN-III-P1-1.1-PD-2019-1029)
Phase 2 (January – December 2021)

I. SUMMARY OF PHASE 2

The objectives of the present project concern the configuration and exploration from a linguistic perspective of an oral speech corpus of contemporary standard Romanian so as to develop new methods of acoustic analysis orientated towards modernizing phonetic research in Romania. The second phase of the project, which took place between January – December 2021, involved corpus development. As such, according to the contract, the following activities have been carried out:

- *high quality audio recordings, spontaneous speech recording, controlled speech recording* – I have recorded all of the experiments (controlled and spontaneous speech) in the Phonetic Laboratory within the Romanian Academy Institute of Linguistics "Iorgu Iordan – Al. Rosetti". Keeping in line with the scientific proposal, all of the 12 participants are representative of the southern dialect on wich the standard language is based on. All protocols involving personal data protection have been respected. In terms of the controlled speech experiment, for vowel analysis and VOT (voice onset time) measurements pertaining to stop consonants, the repetition number per utterance has been increased from 3 (in accordance with the scientific proposal), to 5 repetitions. The motivation behind this choice is based on acquiring larger data, thus contributing to developing an in-depth acoustic study and having a better distribution of the data from a statistical perspective. In terms of the spontaneous speech experiment, the interviews were centered around common discussion topics (see *figure 1*), making the data comparable from a linguistic standpoint.

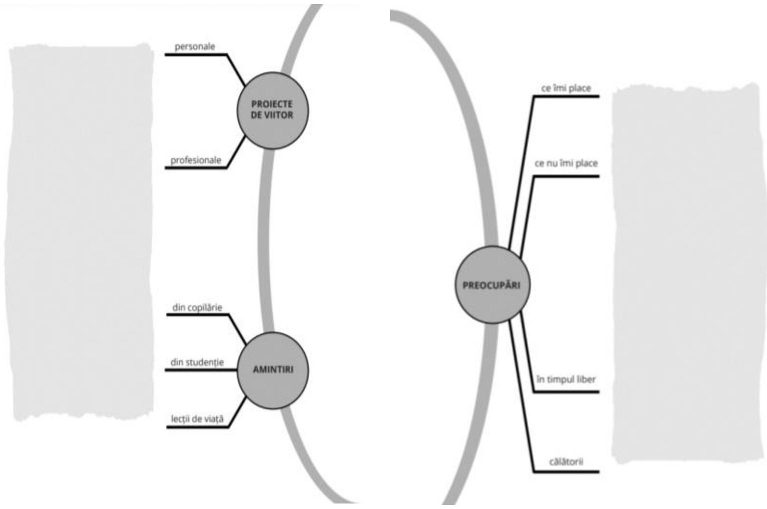


Figure 1. Handout used for eliciting spontaneous speech

- *transcribing and processing the corpus* – the recorded corpus was transcribed and processed. I have transcribed the audio recordings orthographically and phonologically. I addition to the scientific proposal, I opted for introducing various connected speech phenomena (hesitations, truncations, deletions, repetitions, code switching, among others), which in turn can contribute to a faster identification and recovery of the necessary data and offer a better user experience. The phonological transcription was adapted to have the distinction between semivowels and semiconsonants, facilitating, on the one hand, the annotation of the corpus as well as the phoneme – grapheme correspondence, and, on the other hand, contributing to an in-depts phonotactic analysis of spoken Romanian.

○ *analysis in Praat, Excel and R, data visualization* – I have used different software for the quantitative and qualitative analysis of the recorded and transcribed material. Praat was used in different stages of corpus development, more precisely for recording, transcribing, annotating and segmenting. Likewise, the acoustic analyses carried out on the controlled as well as the monologue discourse were done in Praat. The main acoustic measurements were related to duration, fundamental frequency (f0), and the first two formants (F1 and F2), for example:

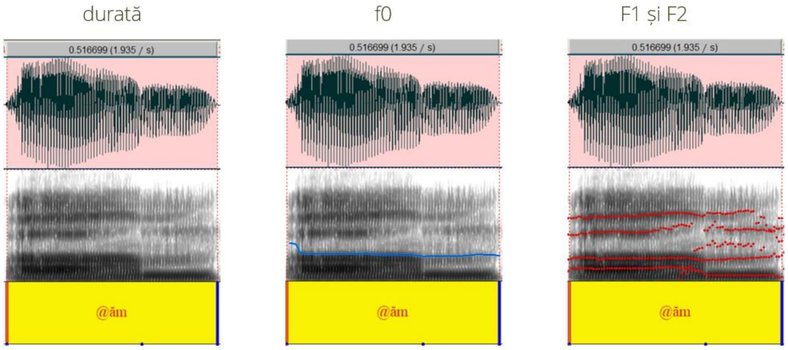
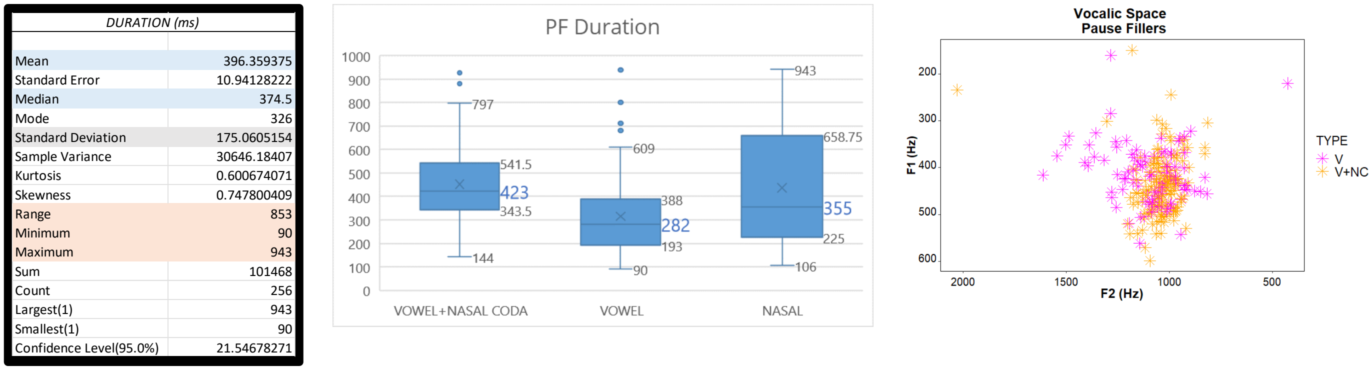


Figure 2. Analysis in Praat

For the exploratory data statistics, I opted for *Excel* and the *Kutools for Excel* add-on. In terms of data visualization, I used *R* and the *RStudio* interface. As an example, for generating the vocalic spaces (developed through extracting F1 and F2 frequencies), I used the *ggplot2* library and wrote the code based on the processed material.

```
library (ggplot2)
setwd ("D:/PAUSE_FILLERS")
data <-  read.table ("type2.txt", sep="\t", header=TRUE)
culori <- c( "magenta", "orange")
ggplot (data, aes (x = F2, y = F1, col = TYPE, shape = TYPE))+ geom_point(size = 5, shape = 8)+ scale_color_manual
(values = culori) + scale_x_reverse()+ scale_y_reverse()+ labs(x = "F2 (Hz)", y = "F1 (Hz)")+ ggtitle("Vocalic Space \n
Pause Fillers")+ theme(panel.background = element_rect(fill = "white", colour = "grey50"))+ theme(panel.border =
element_rect(linetype = 1, color = "black", fill = NA))+ theme(axis.text.x  = element_text(vjust = 0.1, size=15, color =
"black"), axis.title.x = element_text(face="bold", size = 17))+ theme(axis.text.y  = element_text( vjust = 0.1, size=15,
color = "black"), axis.title.y = element_text(face="bold", size = 16))+ theme(plot.title = element_text(size = 19, hjust =
0.5, face = "bold")) + theme(legend.text = element_text(colour="black", size = 17)) + theme(legend.title =
element_text(colour="black", size = 17))+ theme(legend.key = element_rect(fill = "white", colour = "white"))
```



DURATION (ms)	
Mean	396.359375
Standard Error	10.94128222
Median	374.5
Mode	326
Standard Deviation	175.0605154
Sample Variance	30646.18407
Kurtosis	0.600674071
Skewness	0.747800409
Range	853
Minimum	90
Maximum	943
Sum	101468
Count	256
Largest(1)	943
Smallest(1)	90
Confidence Level(95.0%)	21.54678271

Figure 3. Analysis in Excel and R

- *phonotactic description, word-frequency description based on the recorded corpus* – I have explored the transcribed corpus from a quantitative perspective. The analysis was based on the orthographic transcription of the recorded material (where there were marked, alongside other connected speech phenomena, possible elisions within the recorded interviews – this in-depth annotation was not identified in other speech corpora available online, this being an important contribution of our research project towards describing linguistic variation in contemporary standard Romanian), as well as on the phonological transcription adapted to better suit quantitative analyses (for instance, differentiating between semivowels and semiconsonats; distinguishing among vowels surfacing as pause fillers and those within the word). We also focused on analyzing disfluencies in spontaneous speech, such as repetitions (marked by "+" in the annotation), hesitations (marked by "%"), pause fillers (marked by "@"). An in-depth study was carried out on particle fillers, analyzing the frequency of these particles as well as their acoustic properties.
- intermediate release of the speech corpus – I have uploaded for prospective users, samples from the recorded corpus. In this regard, the audio files are correlated with the orthographic and phonologic transcriptions save in TextGrid files. Due to the fact that exploring the corpus requires a minimum familiarity with the Praat interface, we offered the users, in addition to the contractual obligations, handouts highlighting the main functions of Praat. Information related to downloading, visualizing and exploring the corpus are given. Also in addition to the contractual obligations, but keeping in line with the fundamental principles od the scientific proposal, that is stimulating scientific research at the interface between phonetics and phonology, we included in this intermediate launch recordings placing all Romanian consonants in intervocalic position (/VCV/): (1) stops – /apa/ /aba/, /ata/, /ada/, /ak'a/, /ag'a/, /aka/, /aga/; (2) fricatives – /afa/, /ava/, /asa/, /aza/, /aʃa/, /aʒa/, /aha/; (3) affricates – /at͡sa/, /aʧa/, /aʤa/; (4) sonorants – /ama/, /ana/, /ala/, /ara/. All these materials developed for the illustration of Romanian sounds have had numerous applications (for instance, we used the recordings and the associated spectrograms in the practical courses initiated by the project director this year within the Phonetic Laboratory of the hosting Institute; spectrogram reading – see *figure* 4) and are available on the project website.

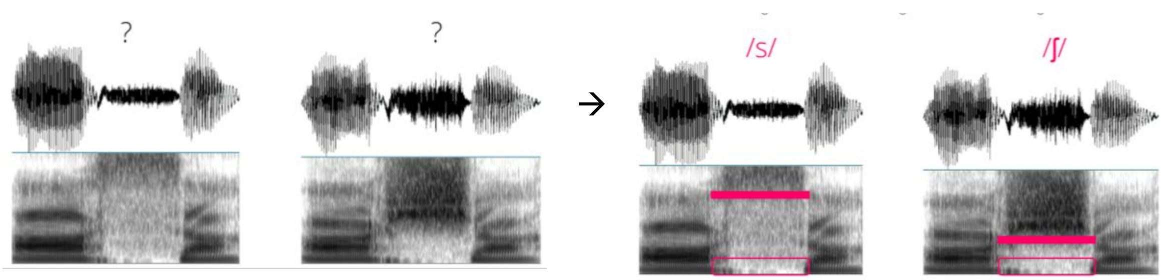


Figure 4. Spectrogram reading (identifying and describing voiced and voiceless fricatives)

- *going on the mobility research stay* – within the postdoctoral research project ROC-lingv (contract number 173/20202), I have carried out a research stay, between 16th October -15th November 2021, at the University in Calabria, Romanian Language and Literature University, Department of Humanities, Italy, coordination Prof. Danilo De Salazar. The papers attesting this research trip can be found on the ROC-lingv project website
- *documentation at the host institute as well as during the research stay abroad* – I have consulted the specialized bibliography related to developing and exploring speech corpora, and during the research stay I also had access to international platforms dedicated to speech corpora.
- *presenting the results within at least one national or international conference* – I have delivered 4 scientific presentations online (see below) regarding the projects results obtain within the second stage, thus exceeding the initial number set out and assumed in the signed contract. I have also delivered 2 presentations as invited speaker at the Humanities Department within the University of Calabria, Italy (see below).
- writing and sending to publication at least one article in a research journal, national or international – I have written and sent for publication one article, titled "Developing linguistic resources for Romanian written and spoken language" (see below).
- *the project website updated* – I have updated the project website in accordance with activities established for the present stage of the project

In terms of project management, I had meetings with the Mentor, prof. univ. dr. Andrei Avram, during which we sketched out the main objectives and intermediate goals for this stage of the project, we set up a research plan, thus ensuring an optimal work flow.

II. SCIENTIFIC AND TECHNICAL DESCRIPTION

The originality of our postdoctoral research resides in designing and exploring high quality oral corpus of contemporary standard Romanian. The corpus will be freely available online, designed to follow the EU recommendations, in agreement with the GDPR norm. By making the corpus available to the general audience, one of the main scientific contributions of the ROC-lingv project is to provide modern resources for the study of Romanian, thus allowing linguistic comparisons with other Romance languages and opening new paths of research

I have presented 4 oral presentations, as single author, at international scientific manifestations:
(1) **Niculescu**, O. 2021. „Constructing an open-source speech corpus of contemporary standard Romanian: outline and preliminary remarks", online presentation at *Phonetics and Phonology: real-world applications (PaPE 2021)*, within the Workshop *From Speech Technology to Big Data Phonetics and Phonology: a win-win paradigm*, 21 – 23 June 2021, Barcelona, Spain.
(2) **Niculescu**, O. 2021. „Recent tools for teaching and conducting research at the interface between phonetics and phonology in contemporary standard Romanian", online presentation at *Al*

21-lea Colocviu Internațional al Departamentului de Lingvistică: Orientări actuale în lingvistica teoretică și aplicată, 19 – 20 November 2021, Bucharest, Romania.

(3) **Niculescu**, O. 2021. „A preliminary acoustic study on pause fillers in Romanian monologue speech. Evidence from a recent developed speech corpus”, online presentation at the Annual International Conference of the Faculty of Foreign Languages and Literature (FLLS 2021), 26 – 27 November 2021, Bucharest, Romania.

(4) **Niculescu**, O. 2021. „Developing linguistic resources for Romanian written and spoken language”, online presentation at *ConsILR 2021* (The 16th edition of The International Conference on Linguistic Resources and Tools for Natural Language Processing), 13 – 14 December 2021, Iași, Romania.

I have also delivered 2 presentations as invited speaker at University of Calabria, Department of Humanities:

(5) **Niculescu**, O. 2021. „Introduction to PRAAT: Tips and Tricks for Doing Phonetics by Computer (Part 1)”, presentation as invited speaker, University of Calabria, 2nd November 2021, Calabria, Italy.

(6) **Niculescu**, O. 2021. „Introduction to PRAAT: Tips and Tricks for Doing Phonetics by Computer (Part 2)”, presentation as invited speaker, University of Calabria, 4th November 2021, Calabria, Italy.

The contribution of the project is acknowledged in each presentation. All contributions are realized in an internationally-spoken language.

I have written and sent for publication one article:

Niculescu, O. 2021. „Developing linguistic resources for Romanian written and spoken language”, in Proceedings of the 16th International Conference “Linguistic Resources and Tools for Processing the Romanian Language”, ConsILR 2021, Iași, Romania.

In conclusion, the project “*Acquiring and exploring an oral contemporary spoken Romanian corpus for linguistic purpose*” (ROC-lingv), coordinated by dr. Oana NICULESCU within the Romanian Academy Institute of Linguistics “Iorgu Iordan – Al. Rosetti”, having as Mentor prof. univ. dr. Andrei AVRAM, has fulfilled all obligations, activities and tasks covered by this second stage of work, often even surpassing the contractual provisions.

Director Proiect,
NICULESCU OANA