ROMANIAN CORPUS FOR SPEECH-TO-TEXT ALIGNMENT

ANCA – DIANA BIBIRI[1], DAN CRISTEA[2,3], LAURA PISTOL[2,3],
LIVIU – ANDREI SCUTELNICU[2,3], ADRIAN TURCULEȚ[1]

[1] *"Al. I. Cuza" University, Department of Interdisciplinary Research in Social-Human Sciences, Iaşi – Romania*
[2] *"Al. I. Cuza" University, Faculty of Computer Science, Iaşi – Romania*
[3] *Institute of Computer Science, Romanian Academy, Iaşi – Romania*

anca.bibiri@gmail.com, {dcristea, laura.pistol, liviu.scutelnicu}@info.uaic.ro, aturcu@uaic.ro

Abstract

In this paper we present the methodology employed in the creation of an aligned speech-to-text Romanian Corpus. The corpus uses recordings from the AMPER-ROM and AMPRom projects as well as ad-hoc recordings of continuous speech. The protocol for speech recording and labelling, as well as the manual annotation procedure, are described. The corpus is intended to be used for training a speech segmentation module and an automatic speech-to-text aligner module.

Keywords: Corpus, speech-to-text, alignment, PRAAT

1. Introduction

Since the early days of intonation research, automatic transcription of the intonation in speech corpora has been on the wish list of many researchers in phonetics, linguistics, and discourse analysis. For several decades, linguistics has gathered a great amount of audio material to study the aspect of spoken language. Unfortunately, some of the recordings have different dialectal signals/marks, for example, background noise, different phonetic intonation, differences in time of intonation and voice changing, etc.

Alignment of the phonemes and text is the first stage of data processing necessary to provide useable training data for many phoneme-to-text conversion systems, including the most successful symbolic rule-based systems and most neural network systems (Bullinaria, 2011).

A common requirement in speech technology is to align two different symbolic representations of the same linguistic message, for instance, phonemes with letters (Damper et al., 2005). As dictionaries become even bigger, manual alignment becomes less and less tenable, yet automatic alignment is a hard problem for a language like Romanian.

In this paper we describe a methodology for building an aligned speech-to-text corpus for Romanian. The investigation has as goal to set the principles of acquiring a significant corpus of signal-text aligned recordings, to be used for training a speech segmenter and a speech-to-text aligner module. By exploiting already existent continuous speech tracks, doubled by their textual transcriptions, an automatic aligner could be used to fabric a large corpus of speeches aligned to their textual transcription, creating thus the prerequisite for training a speech recognition system for Romanian. Other applications of speech-to-text alignment systems are

in fields, such as multimedia indexing, training of large vocabularies for speech recognition, health-related research, etc.

2. Corpora

2.1. AMPER-ROM[ANIA]

L'Atlas Multimédia Prosodique de l'Espace Roman (AMPER) is a last generation atlas which combines principles of geolinguistics with techniques of instrumental phonetics and those of informatics. The atlas is conceived as an interactive database bringing together data collection and acoustic analysis concerning prosodic features of linguistic varieties specific to the Romance languages.

The Romanian Multimedia Prosodic Atlas (AMPRom) is the first prosodic atlas which aims to present the main intonation patterns of the Romanian language varieties identified both at the level of the diatopic variants of the standard language and at the level of the dialect variants.

During the prosodic dialectal investigations, two questionnaires are used: AMPER-ROM[ÂNIA] and AMPRom. The first questionnaire consists of a series of statements (45 sentences) established by morpho-syntactic and phonetic criteria and are formed of: *declarative- affirmative* and *declarative-negative* sentences and total *interogative-affirmative* and *interogative-negative* sentences, having the syntactic structure SVO (subject – verb – object). The S and O receive, in turns, adjective and/or prepositional determinants; the nouns and adjectives that are used in the utterances are trisyllabic oxitones (the last syllable of the word is stressed), paroxitones (the penultimate syllable of the word is stressed) and proparoxitones (the antepenultimate syllable of the word is stressed). Since in the Romanian language the negation usually receives the stress of the phrase, the negative-declarative and interrogative-negative sentences were also introduced in the questionnaire.

The occurrences of the words are at the right and at the left of the verb for capturing all the prosodic indices (S – subject, V – verb, O – object, Adj – adjective – with the mention that the subject is interchangeable with the object):

[S + V + O / S + Adj/ + V + O / S + V + O + Adj / S + S + V + O / S + V + O + S]

AMPER-ROM questionnaire (sequence) (Each sentence is labeled with a unic code in order to identify the sentence when the acoustic analysis is made: *bwt*, *dwk*, *fwt*, *gwt*, *kwt*, *pwt*, *swk*, *twg*, *twk*, *zwk.*):

twk *Nevasta vede un căpitan.*/ *The wife sees a captain.*
kwt *Un căpitan vede nevasta.*/ *A captain sees the wife.*

dwk *Nevasta tinerea vede un căpitan.*/ *The young wife sees a captain.*
gwt *Un căpitan elegant vede nevasta.*/ *An elegant captain sees the wife.*

swk *Nevasta frumoasă vede un căpitan.*/ *The beautiful wife sees a captain.*
pwt *Pasărea vede nevasta.*/ *The bird sees the wife.*

zwk *Nevasta harnică vede un căpitan.*/ *The hardworking wife sees a captain.*

bwt *Pasărea papagal vede nevasta.*/ *The parrot bird sees the wife.*

twg *Nevasta vede un căpitan elegant.*/ *The wife sees an elegant captain.*
fwt *Pasărea frumoasă vede nevasta.*/ *The beautiful bird sees the wife.*

There are in AMPER-ROM questionnaire sentences with broad focus, as in the following examples. The labels of the sentences represent: *twkae1* – the declarative affirmative sentence with the focus on the first element – subject; *twkie2* – the interrogative affirmative sentence were the object is stressed; *twknev* – the declarative negative sentence with focus on the verb.

twkae1 ***Nevasta*** *vede un căpitan./* ***The wife*** *sees a captain.*
twkie2 *Nevasta vede* ***un căpitan****?/ The wife sees* ***a captain****?*
twknev *Nevasta* ***nu vede*** *un căpitan./ The wife* ***does not see*** *a captain.*

2.2. AMPRom

In order to capture a larger number of Romanian intonation patterns in their territorial distribution, a second questionnaire includes other statements, simpler (with not so many formal constraints) to facilitate the contact with the subjects and to prepare them for the fixed questionnaire. This includes about 100 sentences and has two variants: short version (compulsory, with 84 sentences) and extended version (optional, having 111 sentences), the latter is applied only at some points of inquiry.
Types of syntactic structures that make up the AMPRom questionnaire:

- VO structures (with inclusive subject): 1a: *L-ai văzut pe Ion?/ Have [you] seen John?* 3a: *Ai văzut fetele?/ Have [you] seen the girls?*
- Structures pursuing the relation between the word order and prosody: (1) 1b: *Pe Ion l-ai văzut?/ John was that you have seen?* 3b: *Fetele le-ai văzut?/ Girls were that you have seen?*
- VS/SV Structures: 25a: *Vine Ion./ There comes John.* 25b: *Ion vine./John is coming.* 28a: *Cine vine?/ Who is coming?* 28b: *Ion vine./John is coming.*
- Structures with double negation elements both in the question and in the answer: (26): *Nu vine nime(ni) la noi?/ There comes There comes nobody/none to us?* (30): *N-a venit nime(ni) la noi./Nobody/none came to us.*
- Structures in which modulators are used (adverbs of manner and semi-adverbs – *sure, precisely, certainly, immediately, surely, maybe, whether, really* or even modal verbs – *I think, it might*): 20b: *Chiar vine Ion?/ Really, is John coming?* 21a: *Sigur/Precis (că) vine/ Sure/precisely he is coming.* 23c: *Cred că vine./ I think he is coming.*
- Structures containing different types of questions: partial, alternative, confirmation: 56a: *Cât e ceasul?/ What time is it?* 41: *Vii ori nu vii?/ Are you coming or not?* 55b: *Pleci mâine la Iaşi, nu-i aşa?/ You are going tomorrow to Iaşi, aren't you?*
- Structures containing vocative addressing and calling: 40: *Ion (Ioane), dă-mi un măr (te rog)! / Ion (John), give me an apple (please)!* 35a: *Ana!/ Ann!,* 35b: *Maria!/ Mary!,*
- Structures that require an intonation of continuity (in suspension): 49: *– Apucă-te/Ia şi-nvaţă, că de nu.../ Start/ Put yourself at work/to learn, or else...*
- Exclamatory structures: 84: *Ce batic frumos ai!/ What a beautiful scarf [you] have!*
- Structures on intercalation prosody: 74a: *Tata mi-a zis: Du-te repede şi cheam-o pe soră-ta! / My father said, 'Go quickly and call your sister'!* 74b: *Du-te repede şi cheam-o pe soră-ta! mi-a zis tata. / Go quickly and call your sister! my father said.*
- Structures containing enumerations: 66: *Am fost la piaţă/târg şi am cumpărat: roşii, ceapă, morcov şi ardei./ I was at the market/fair and bought tomatoes, onions, carrots and peppers.*

- Structures containing a sequence of short sentences: 79: *De dimineaţă m-am trezit, am pregătit micul dejun şi apoi am plecat la serviciu./ This morning I woke up, I made breakfast and then I went to work.*
- Sentences with the same structure (V) for the affirmative, interrogative and imperative mood: 80: *Aşteaptă./[He/she]waits.* 81: *Aşteaptă?/Does [he/she] wait?* 82: *Aşteaptă!/Wait!/ Aşteaptă-mă!/Wait for me!*
- Structures with a focus on different constituents 4a: Pe **Vasile** l-ai văzut ?/Was **Basil** that you saw? 4b L-ai văzut pe **Vasile**?/ Did you see **Basil**?; 58: Bei **vin**?/Are you drinking **wine**?
- Structures with a successive focus on constituents 64: ***Mănânci*** *peşte?/ Are you* ***eating*** *fish?* 65a: *Mănânci* ***peşte****?/ Are you eating* ***fish****?*
- Affective structures: 56f: *E/îi amiază? / Is it/ It's noon? It's already noon?* 59: ***Bei vin?/Are you drinking wine?***
- The extended form of the questionnaire contains other type of syntactic structures:
- Structures pursuing the prosody of idioms and phrases: 89 a, b, c…: *da de unde!/what? no way!; nu mai spune!/ yah, do not say!; ce folos?/ so, what?; nici vorba/pomeneala!/no way!/not at all!; cum/unde să facă ea aşa ceva?/what/how did she do that?; da mai ştii?/that could be?;ei şi?/so, what?.*
- Structures containing greetings and politeness: 91: *Bună ziua! Good afternoon!*; 97: *Poftim/There you go!/ Na!/Here! – Mulţumesc/mulţam!/ Thank you/Thanks!*
- Structures that use adverbs and adverbial phrases to strengthen the assertion and negation: 104: *Da, sigur/ fireşte/ negreşit!/ Yes, sure/ surely/ no doubt!* 105: *Nicidecum!/No way! Niciodată/Never! Nici în ruptul capului!/On no account!*
- Imprecations: 107 a, b, c…: *Arde-l-ar focu' să-l ardă!/ May he burn in hell! Lua-l-ar naiba/dracu să-l ia!/ The hell/the devil with him! Fir-ar/fi-o-ar a dracului!/ Damn it/Damn with it! – Du-te dracului/la dracu/la satana!/ Go to the devil/to Satan!*

The statements are recorded at least three times and are obtained through indirect questions and by verbal and non-verbal implications (facial expressions, gestures) to the context, and/or forming some speech situations during the continuous dialogue between the investigator and the informant.

In rural areas, two indigenous subjects are used, representative for the local speech, with elementary education, middle-aged, who speak natural under the conditions of the investigation. In urban areas the surveys are twofold: besides the informants belonging to low and/or middle class with influences of the local dialect, there are used subjects with higher education, speaking a cultivated language.

2.3. The IIT corpus

The IIT continuous speech corpus consists of recordings, summing up 45 minutes of continuous speech, uttered in an office environment and following a standard voice recording procedure, by three female speakers who currently speak Romanian standard language, aged between 33 and 50, having no pathological disorder and originated from the geographical area of North-Western Romania (the Iași district). The recordings were single channelled with a sampling frequency of 22050 Hz and 16 bit resolution. The sentences chosen for recordings are paragraphs from "Amintiri din copilărie" (*Childhood Memories*), by the classical Romanian writer Ion Creangă and dialogues from sketches by the Romanian writer and dramatist Ion Luca Caragiale. The choice towards this piece of classical belletrist work was imposed by the necessity for the corpus to be copyright-free.

The size of the IIT database is shown in the following table:

Table 1: Size of the database (Only for the writer Ion Creangă)

sentences	341
vocabulary size	2000
words (occurrences)	6505
words per sentence	19.07

3. Notation of sounds, phonemes, graphemes

In the following, by *sound* we mean a segment of a speech track, as it is heard by a human or is recorded by a machine. A sound, in general, is characterised by steady physical parameters (amplitude, frequency) and corresponds to a letter in an alphabetic transcription. There is a huge variance of sounds corresponding to the same letter, depending on the articulatory and the co-articulations conditions of the sounds and to other factors, such as the context of communication and the speaker (sex, age, tonality, momentary physical and psychological state).

A *phoneme* is the conceptualization of a sound. The Romanian language has 31 phonemes. As such, one cannot say that phonemes are recorded. Only sounds can be recorded, but out of them, phonemes are deciphered (interpreted) and, accordingly, noted. In the real world, a phoneme does not exist, but we can say "this sound records the phoneme *a*". The phonemes are noted in the International Phonetic Alphabet – IPA (see below).

The *speech-to-text* alignment conventions are based on the mappings between the two planes of expression of language: the concrete plane (of the substance of the language), populated with sounds, and the abstract plane (of the form, the linguistic plane), where phonemes coexist. These two planes are both doubled by two levels of expression: phonic and graphic (as suggested by Figure 1).

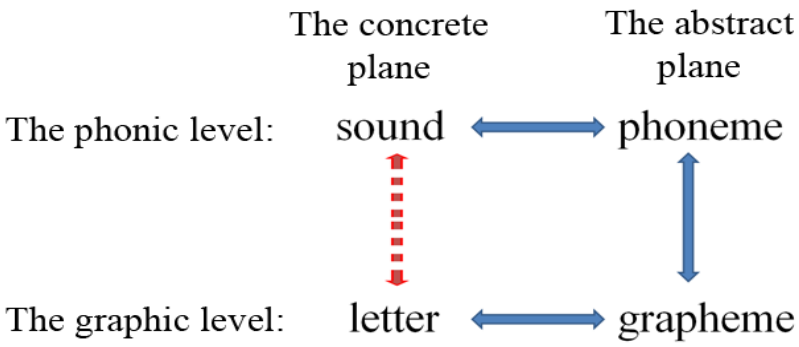

Figure 1: The speech to text correspondence

Although a phonemic language (sounds as they are transcribed), Romanian has some particularities:

- the sounds z̲ and s̲ in *dezbat* vs. *desfac*, or *răzbate* vs. *răsplati* have as a variant /S/: /deSbat/, /rəSbate/; *îmbrac* and *învăţ* have as a variant /N/: /ɨNbrak/, /ɨNvətz/; as a result of neutralization of the opposed /z/ and /s/, respectively, between /n/ and /m/ in such examples it is noticed the occurrences of the archiphonemes /S/ and /N/.

- the use of morphematic principle in order to maintain the formal identity of the words, especially when one speaks about the alternation of the diphthongs *oa* and *ua*, respectively *ea* si *ia*: *oa ~ o*: *oameni – om, toată – tot , oală – ol; ua ~ u: băcăuan – Bacău, flăcăuaş – flăcău*, and in the case of some neologisms: acuarelă, scuar; *ea ~ e*: *teamă – tem, cheamă – chem, ceas – cesuleţ, ea – ele*; *ia ~ ie*: *iarnă – ierni, piatră – pietre* or in the situation when there is no alternative: *chiar, ghiaur*;

- the morphematic principle is rarely used to differentiate morphemes: *aceea(şi)* vs. *aceiaşi, ea* vs. *ia*;

- it is maintained (totally or partially) the etymological spelling: *eu, el, ei, ele, eram*; *absent, lied, watt, subţire, fotbal; alură, bleu*; the most typical case is that of loans from English: *computer, laptop, site, whisky, weekend*.

Romanian spelling includes graphemes created using diacritical marks (because of the lack of specific letters in the Latin alphabet: ă, â, î, ş, ţ) as well as polyvalent, compound graphemes having different contextual values.

There are polyvalent vocalic graphemes <e>, <i>, <o>, <u>, noting both the vowels /e/, /i/, /o/, /u/, and the corresponding semivowels /e̯/, /i̯/, /o̯/, /u̯/; also the sequence of a vowel + a dependent semivowel: <e> = [i̯e] in *eu, eram, vie*; <i> = [i̯i] in *cais, fiinţă, oişte* or [ii̯] in *academia, ia* ‘bluză’; <o> = [u̯o] in *fior*, [ou̯] in *merituos*; <u> = [u̯u] in *aur* or [uu̯] in *luând*. In some cases, according to the morphematic principle, the graphemes <e>, <o> also note semivowels <i>, <u>: *aceea, ea, oameni, vioară*.

The consonantic graphemes <c>, <g>, <k>, <n>, <x> have double values depending on the context where they occur: /k/ and /tʃ/, /g/ and /ɟ/, /k/ and /c/, /n/ and /N/, /ks/ and /gz/ in *car* and *cer*, *gară* and *ger*, *kaliu* and *kaki*, *nas* and *învăţ*, *aks* and *exemplu*.

There are graphemes compound of two or three letters: <ce>, <ci>, <ge>, <gi>, <ch>, <gh>, <che>, <chi>, <ghe>, <ghi>: *ceas, arici, geam, ungi, chem, ghem, cheamă, gheaţă, ochi, unchi, unghi*.

The description of the phonemic system of the Romanian language has several interpretations, with different numbers of phonemes, depending on the authors’ theoretical and methodological assumptions. The Romanian linguist E. Petrovici (1956) proposes in his phonemic theory the largest number of phonemes: 5/7 vowels and 72 consonants, and E. Vasiliu (1965) – the smallest number of phonemes: 7 vowels, 20 consonants, and one special phoneme called ’syllabic juncture’.

For our corpus we propose a simple phonemic system, which best corresponds to Romanian writing, in accordance with the Latin alphabet.

This phonemic system (Turculeț, 1999) is made up of 7 vowels: /e/, /i/, /a/, /ə/, /ɨ/, /o/, /u/, 4 semivowels /e̯/, /i̯/, /o̯/, /u̯/ and 20 consonants ([c], [ɟ] are considered allophones of the phonemes /k, g/) – see Table 2.

Table 2: Symbols for consonants

Place→ **↓Manner**	**Bilabial**	**Labio-dental**	**Dental-alveolar**	**Alveolar**	**Alveolo-palatal**	**Velar**	**Glottal**
Plosive	/p/ /b/		/t/ /d/			/k/ /g/	
Nasal plosive	/m/		/n/				
Fricative		/f/ /v/	/s/ /z/		/ʃ/ /ʒ/		/h/
Affricate			/ʦ/		/ʧ/ /ʤ/		
Lateral				/l/			
Trill				/r/			

The reduced vowel i̯, asyllabic and voiceless, specific to the Romanian language called' final, asyllabic, post-consonant i' such as in [lupĭ], [potzĭ] (it occurs rarely within a compound word, at the morpheme limit [orĭkɨnd], [kɨtzĭva]) as a variant of semivowel /i̯/. Thus, the phonetic label is [ĭ] and the phonematic one as /i̯/ (it occurs after a consonant in the final position and between two consonants in medial position).

The back rounded vowels [ö] and [ü] originated in some French and German loans can be considered as situated at the Romanian phonetic and phonemic periphery: <alură> [alürə], <bleu> [blö], <röntgen/roentgen> [röntɟen] or [röntgən]. They are realised usually as the diphthongs [i̯u], respectively [e̯o].
Regarding the correspondence between phonemes and graphemes we propose some simple solutions according to the combinations of Romanian letters used in writing. They concern the evaluation of compound graphemes (see *supra*). The compound graphemes from the following examples <*ceas*> [ʧas], <*arici*> [aritʃ], <*geam*> [dʒam], <*ungi*> [undʒ] are reduced to simple graphemes <c>, <g>, followed by the 'latent' phonemes /e̯/ and /i̯/ (possible solution proposed in generative phonology) with the phonemic transcription / ʧe̯as/, / aritʃi̯/, /dʒe̯am/, / undʒi̯/, and the trigraphs <*che*>, <*chi*>, <*ghe*>, <*ghi*> followed by vowels <*a*>, <*o*>, <*u*> or in final position are reduced at <*ch*>, <*gh*> : <*cheamă*> [camə], <*chiar*> [car], <*chior*> [cor], <*chiul*> [cul], <*gheaţă*> [ɟatzə], <*ghiozdan*> [ɟozdan], <*ochi*> [oc], <*unghi*> [uŋɟ], with the phonemic transcription /ce̯amə/, /ci̯ar/, /ci̯or/, /ci̯ul/, /ɟi̯atzə/, /ɟi̯ozdan/, /oci̯/, /uŋɟi̯/.
The compound graphemes are, in fact, the digraphs <*ch*> and <*gh*> as in <*chem*> [cem] /cem/, <*ghem*> [ɟem] /ɟem/, <*cheamă*> [camə] /ce̯amə/, <*gheaţă*> [ɟatzə] /ɟe̯atzə/, <*ochi*> [oc] /oci̯/, <*unchi*> [uŋc] /unci̯/, <*unghi*> [uŋɟ] /unci̯/.

Figure 2 shows an example of a speech-to-text alignment: partial interrogative sentence uttered by a subject from Bucharest (Cristina Dăbuleanu, 49 years old, computer programmer): *Cum te cheamă?* (*What is your name?*).

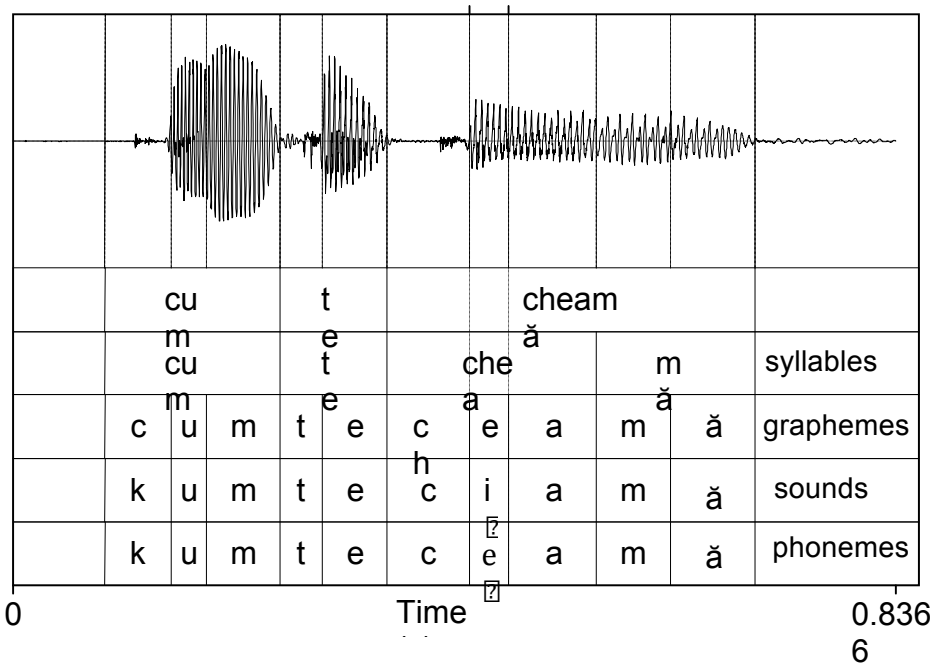

Figure 2: Praat screen in the speech-to-text alignment of the utterance *Cum te cheamă?*

For some loans (most of them from English), there are applied the rules for writing and speaking of foreign language, will be marked with a special sign. The letters/graphemes and the sounds/phonemes will be maintained as they are in the foreign language: <*laptop*> [læptop], <*site* > [sajt].

4. Speech-to-text alignment

The purpose of the manual speech-to-text alignment is to determine with precision the boundaries of sounds belonging to the phonic layer and to align them with letters from the grapheme layer. The task is done by one of the co-authors, having an extensive experience in reading spectrograms and labelling phonemes. By using the graphical interface and listening the audible track in Praat, she identifies the acoustic changes in order to determine the phoneme boundaries. The annotation levels are: utterance, word, syllable, phoneme and grapheme. Table 3 shows the notations used with Praat in the alignment process.

Table 3: 9 tracks revealed by Praat, shown at different moments of time; apart from duration, the first 3 tracks (sound, syllable and word) represent manual annotation, while the other 5 are automatically recorded

Time	Duration	Sound	Syllable	Word	Intensity(dB)	F0 (Hz)	F1	F2	F3 (Hz)
0	0.09	n	ne	nevasta	74	0	390	1768	3096
0.09	0.05	e/\ef	ne	nevasta	73	205	463	1835	3088
0.14	0.06	v	\'1vas	nevasta	70	0	567	1484	2220
0.2	0.12	a::	\'1vas	nevasta	78	237	807	1523	3187
0.32	0.06	s	\'1vas	nevasta	61	0	806	1728	3208
0.38	0.06	t	ta	nevasta	75	0	621	1484	2965
0.44	0.06	a	ta	nevasta	79	228	875	1529	3092
0.5	0.08	v	\'1ve	vede-un	63	0	648	1375	2780
0.58	0.06	e	\'1ve	vede-un	77	230	497	2252	2927
0.64	0.07	d	di-un	vede-un	70	0	371	2291	2958
0.71	0.08	i\nvu\~^	di-un	vede-un	70	236	399	1087	2776
0.79	0.04	\ng	di-un	vede-un	71	0	452	1010	2593
0.83	0.06	k	c\sw	c\swpitan	63	0	257	1634	2912
0.89	0.04	\sw	c\sw	c\swpitan	76	223	527	1528	2827
0.93	0.09	p	pi	c\swpitan	53	0	464	1903	3146
1.02	0.04	i	pi	c\swpitan	69	212	389	2504	3353
1.06	0.1	t	\'1tan	c\swpitan	53	0	301	2541	3395
1.16	0.09	a\~^\~v:	\'1tan	c\swpitan	66	146	942	1746	3229
1.25	0.06	n	\'1tan	c\swpitan	62	0	1039	3137	3568

PRAAT is a flexible tool for the analysis of acoustic speech signals. It offers a wide range of standard and non-standard procedures, including spectrographic analysis, articulatory synthesis and neural network. Speech segmentation is the process of identification of boundaries between words, syllables and phonemes. Performed manually, this process attaches a label to each segment. For example, after we have finished segmenting the words and labelled them, follows the segmentation of the syllables of the structure and, finally, those of the compound sounds. The steps in the analysis of a speech waveform are as follows:

- The script reads sound files (.wav format – *Waveform Audio File Format*) from a user-specified folder;
- Then create a TextGrid (which consists of a number of tiers – an interval tier is a connected sequence of labelled intervals, with boundaries in between);
- Selecting both .wav and Text Grid files it opens a window spectrogram in which the annotation is made manually: 3 tiers are open in order to annotate words, syllables and phonemes;
- Once the speech signal is segmented and labelled, by pushing the run button a text file is generated in output, including different parameters: the fundamental frequencies (F0, in the three points of a vowel – F1, F2 and F3), the duration and the intensity of the acoustic signal.

For the speech-to-text alignment of the corpus, the supra-segmental features of the utterance are also taken in consideration: the stress, the intonation and the break indices (as indicated by punctuation marks). A more appropriate rendering is that used in ToBI[1] – a framework for developing community-wide conventions for transcribing the intonation and prosodic structures of spoken utterances in a language variety. A ToBI framework system for a language variety is grounded on the intonation system and the relationship between intonation and the prosodic structures of the language.

[1] Tones and Break Indices: http://www.cs.indiana.edu/~port/teach/306/tobi.summary.html

5. Conclusions

In this paper we presented a methodology of manual annotation of an aligned speech-to-text corpus for Romanian, and the phonetic peculiarities of this language. The intention is to use this corpus to train a speech segmentation and aligner program (let's call it a SEG-ALI module) that would be able to detect the boundaries of sounds in correlation with a text track where the textual transcription is noted. Different parameters of the speech signal, some of them having been suggested in this paper by presenting the processing capabilities of the Praat system, will be exploited by a learning system that will finally train the SEG-ALI module. A top-down strategy will, most probably, be employed for this purpose, by searching first the pauses in the sound track and aligning them with the boundaries between sentences and words and using more high level features to detect phonemes boundaries in between pauses of the continuous speech.

Once such a SEG-ALI module is obtained, it could be used to segment and align automatically a very large corpus of parallel tracks containing human produced continuous speech and their textual transcription. In the long run, the intention is to acquire a large corpus of aligned speech-to-text records that will be used in training a speech recognition system for Romanian. Knowing the high costs encumbered by manual segmentation of the voice track and its alignment against the text track, our hope is to arrive at a very good performance of the SEG-ALI module that would permit the automatic acquisition of a very large corpus in a short time and with reduced costs.
We do not neglect also the possibility to use a boot-strapping strategy in acquiring a high quality aligned corpus: use the manually annotated corpus as a core corpus on which a beta version (v0) of a SEG-ALI module is first trained. Use then this SEG-ALI-v0 to segment&align a larger corpus, and then involve specialised humans to correct it. This activity is supposed to take less time than building it from scratch and also cost less. Once finished, use this larger corpus to retrain the SEG-ALI module to a new and enhanced version – v1, and so on.

References

AMPER – Atlas Multimédia Prosodiques de l'Espace Roman, http://w3.u-grenoble3.fr/dialecto/AMPER/amper.html

AMPROM – http://amprom.uaic.ro/

Handbook of the International Phonetic Alphabet. A Guide to the use of the International Phonetic Alphabet (1999). Cambridge University Press.

Boersma, P., Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.42, retrieved 8 February 2013 from http://www.praat.org/

Bullinaria, John A. (2011). Text to Phoneme Alignment and Mapping for Speech Technology: A Neural Networks Approach, IJCNN, IEEE, 625-632.

Damper, R. I., Marchand, Y., Marsters, J.-D. S., Bazin, A. I. (2005). Aligning text and phonemes for speech technology applications using an EM-like algorithm. *International Journal of Speech Technology*, no. 8, 149-162.

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states, *Speech Communication*, no. 51, 352-368.

Petrovici, E. (1956). Sistemul fonematic al limbii române (*The phonemic system of the Romanian language*), in Studii şi Cercetări Lingvistice, VII :1-2, 7-21.

Turculeţ, A. (1999). Introducere în fonetica generală şi românească (*Introduction to general and Romanian phonetics*), Demiurg Editorial House, Iaşi.

Vasiliu, E. (1965). Fonologia limbii române (*The phonology of the Romanian language*), Editura Ştiinţifică, Bucharest.