Building a Representative Audio Base of Syllables for Romanian Language

Ștefan - Stelian Diaconescu,
Monica - Mihaela Rizea, Mihaela Ionescu,
Andrei Mincă, Liviu Dorobanțu, Ștefan Fulea,
Monica Rădulescu
Research & Development Department
SOFTWIN
Bucharest, Romania

Horia Cucu, Dragoș Burileanu
Speech and Dialogue Research Laboratory
University Politehnica of Bucharest
Bucharest, Romania

***Abstract*—The aim of this work is to provide some insights regarding the effort of building a representative and wide coverage audio base of syllables for Romanian. The audio base comprises audio recordings of syllables extracted from the following types of syllable embedding: isolated-syllable, isolated-word and continuous speech. The list of syllables has been computed over the syllabified form of single-word inflected forms. The inflected forms were generated using a general rule-based system for normal and phonetic inflection having at its core the GRAALAN (GRAmmar Abstract LANguage) metalanguage (designed for linguistic knowledge description). In addition, the word-position of a syllable was accounted for when planning the audio recordings.**

Keywords—language resources; audio base of syllables; very large vocabulary; isolated-syllable pronunciation; isolated-word pronunciation; continuous speech;

I. INTRODUCTION

It is generally known that the phonetic databases reported for Romanian contain a reduced number of entries and most of them are manually obtained. One of the most noticeable resource for Romanian is given in [1], a corpus of 11,819 manually-corrected syllabified words transformed in their phonetic forms. The necessity of automatic phonetic transcriptions and automatic syllabifications using internationally approved standard is constantly growing in the era of artificial intelligence and speech recognition systems.

In this paper, we focus on large vocabularies, of approximately 100.000 lemmas and 1.250.000 inflected forms, which account for approximately 12,000 different syllables (with stress marker variants included). Such resources are at the core in any application or system that engages human-machine interfaces.

Moreover, in designing a human-machine speech interface the acoustic model is a central proccesing step and presents different challenges depending on the chosen sound-unit (phoneme, phoneme bigram and/or trigram, syllable, word etc.) [2][3][4]. For example, in one effort for Romanian, in the training strategy for Hidden Markov Models (HMM) the approach that led to the best results on automatic speech recognition (ASR) is the use of the *triphone* sound-unit (phoneme trigram) [5]. Considering that the nature of syllable is of such that in the majority of cases it consists of more than one phoneme and much more importantly the syllable involves a particular speech effort, one can assume that the syllable could provide a promising hypothesis for the acoustic model in human-machine speech interface systems [6][7][8]. A syllable-based unit approach for English language has proven the advantage of syllable units that capture the dynamics of the speech and particularly the dynamics within phoneme boundaries and in the transition region [9].

The lack of specific linguistic knowledge and of audio utterances for syllables has been a significant setback for the syllable-based acoustic models, especially for Romanian. Unfortunately, in the case of audio databases the resources are still scarce even for the popular languages [10][11]. This is due to the laborious process of gathering audio utterances: either from isolated-syllable pronunciation or by manually segmenting the continuous speech and tagging the syllable transcription. In the case of linguistic knowledge bases for Romanian the status has improved lately with many efforts involving both language corpora and or rule-based resources [12].

Our priority is to obtain the detailed description of Romanian syllables and to build a qualitative audio base for syllables. We base our research on GRAALAN system, which has the advantage of being applicable for generating large-scale phonetic databases and of being able to generate results with a high level of accuracy.

This paper is organized as follows: Section II presents a GRAALAN view for the concept of *syllable*; in Section III we give the basic information for syllabification and inflection methods in GRAALAN system; in Section IV we present the extraction and listing of syllable situations from a GRAALAN database; Section V explains how we created a minimal set of words, that covers the lists of syllable situations, that is subject to audio-recording in order to form the preliminaries of an audio base of syllables; in Section VI we present the construction of a representative audio base of syllables for Romanian language and Section VII contains a number of statistics and results.

978-1-5090-6497-7/17/$31.00 ©2017 IEEE

II. THE SYLLABLE

The concept of syllable has received various definitions, both as phonetic and phonological views. According to A Dictionary of Linguistics and Phonetics [13], a syllable is *"a unit of pronunciation typically larger than a single sound and smaller than a word"* and consists of three parts: the onset (optional), the nucleus (or centre, which is obligatory) and the coda (optional). According to DEX [14], a syllable is *"a phonetic segment that consists of one or many phonemes, pronounced by a single expiratory effort"*.

In addition, we regard the syllable as a set of phonemes clustered around a stressed vowel. This cluster consists of the following:

- (possibly) a set of vowels and/or consonants preceding the stressed vowel
- a stressed vowel (the central vowel)
- (possibly) a set of vowels and/or consonants following the stressed vowel

In continuous speech, a syllable has a relatively constant duration. The speech has a certain rhythm which generally varies slightly from a syllable to another. This is more evident when reciting a (classical) poem (with constant rhythm and meter, and possibly rhyme). Therefore, although the syllables have a variable number of phonemes (which, in turn, have different durations), generally have a relatively constant duration (within the same word or during the same "utterance"). We call this duration Standard Syllable Duration (SSD).

In this, we will consider a word as a set of one or many syllables. One of the syllables of a word is more stressed than the others, meaning that the central vowel is more stressed than the others and we marked it with *"primary_stress"* - see more in Fig. 1. Some languages have more accurate rules regarding the position of the stressed syllable in a word. For example, in French, the stressed syllable is usually the last syllable in a word, while in Hungarian, the stressed syllable is the first one in a word. In Romanian there is no rule regarding the position of the stressed syllable in a word.

In certain words (usually longer) there is a second stressed syllable and we marked it with *"secondary_stress"*. As a result, a word can have the following types of stress:

- *primary stress*: attached to the central vowel of the stressed syllable
- *secondary stress* (*possibly*): attached to the central vowel of a syllable different from the syllable carrying the primary stress
- *normal stress* (*possibly*): attached to the central vowel of a syllable different from the syllable carrying primary or secondary stress

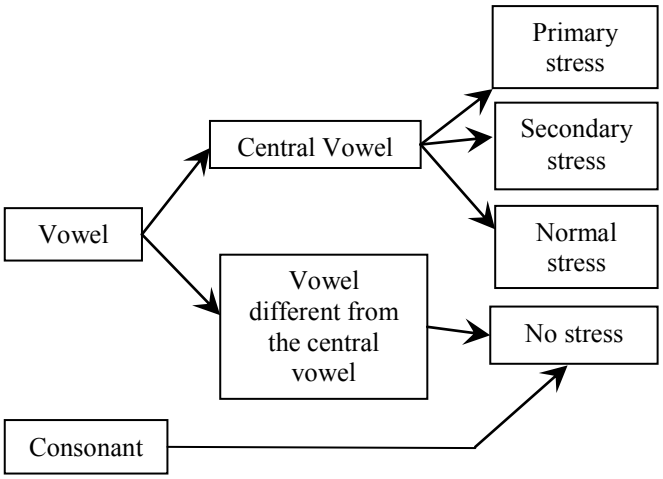

Fig. 1. Types of phoneme stress markers

Moreover, one can highlight that the word-position of a syllable is a defining characteristic. Thus the syllable adjacency can be of the following types:

- left or right *syllable* adjacency (adjacency with another syllable in the same word)
- left or right *word* adjacency (adjacency with another syllable in another word)
- left or right *silence* adjacency

In principle, almost any word may follow any word in a text, hence the (left or right) adjacencies for distinct syllables or distinct syllable pairs would be in large number. Therefore, we could neither obtain phonograms for them, nor could they be processed with reasonable costs. In this regard, we will consider only some of these adjacencies.

III. GRAALAN SYSTEM: INFLECTION AND SYLLABIFICATION

A. The GRAALAN system

The GRAALAN system [15] includes the following components (see Fig. 2):

1. **Theoretical** bases, that consists of notions such as Attribute Value Trees (AVT), Dependency Trees (DT), Generative Dependency Grammar (GDG), Generative Dependency Grammar with Features (GDGF), tetravalent logics.

2. The **GRAALAN** Metalanguage developed based on the theoretical notions [16]. This powerful instrument allows the linguist to formally describe various linguistic sections, such as:

   a. The **Alphabet** of a natural language (including also the phonetic alphabet (through which one can describe the spelling of words))

   b. **Syllabification Rules** (taking into account the euphonic, phonetic, and morphological characteristics).

c. The **Morphology** (morphological categories, morphological category values and their restrictions).

d. The **Lexicon**, including morphemes (prefixes, suffixes etc.), lemmas (with multiple characteristics), multiword expressions (MWEs), and morphological structures (morphological categories with their values, synthetic and analytical entries).

e. **Inflection Rules** describing the operations through which, starting from the lexicon lemmas, one can (automatically) obtain all the synthetic and analytical forms (the entire word paradigm). These rules apply to the forms in normal and phonetic alphabet.

f. **Inflected forms** obtained by applying the inflection rules to the lexicon lemmas. Each inflection form contains its variant in normal and phonetic alphabet, its syllabification (in normal and phonetic alphabet), and the characteristics of the inflection situation.

g. The **Syntax** containing the set of rules that govern the structure of the sentences in a natural language. This is a very complex level, especially for languages with rich inflection and with relatively free word order (such as Romanian).

h. **Correspondences between two natural languages** (between words, multiword expressions, morphological and syntactic structures).

3. **Linguistic instruments (tools)** that help the linguist to work with the GRAALAN metalanguage and to handle a huge volume of (linguistic) knowledge. These tools include: the GRAALAN Compiler, the Lexicon Knowledge Tool (LKT), the Morphological Knowledge Tool (MKT), the Bilingual Dictionary Knowledge Tool (BDKT), and the Linguistic Knowledge Bases Linkage Cheker (LINK).

4. **Linguistic Knowledge Bases** (LKB) containing linguistic data regarding the alphabet, the syllabification rules, the morphology, the lexicon, the inflection rules, the inflected forms, the syntax, and the correspondences between two natural languages. All of these are represented in different formats, mostly in GRAALAN and in XML. There are also pieces of information represented in relational databases and accessed through different linguistic tools.

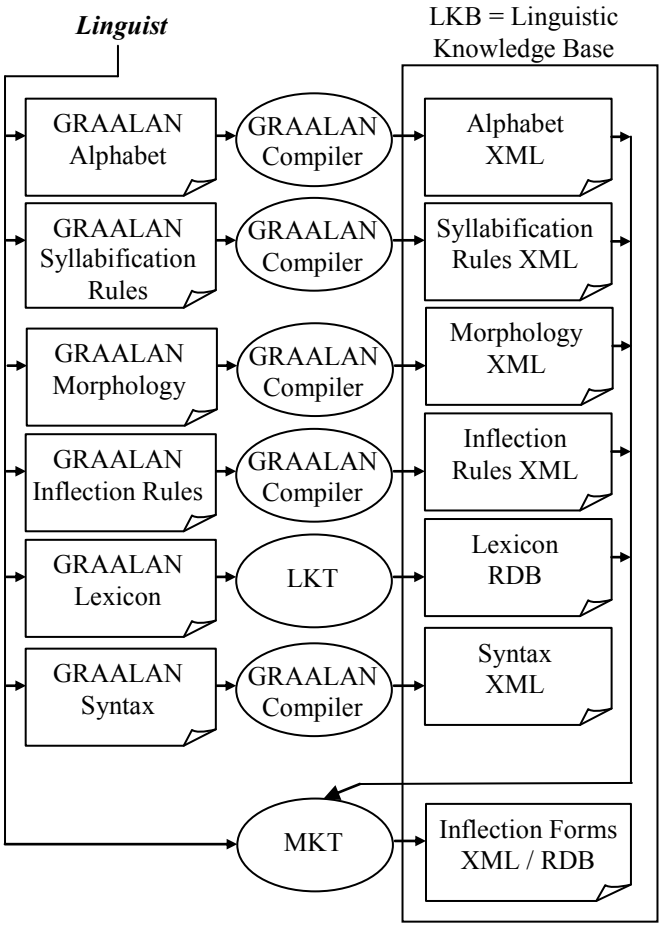

Fig. 2. Example of a GRAALAN system operation flow

5. **Applications.** The Linguistic Knowledge Bases (LKB) are exploited by various studies and applications (see Fig. 3). In this case, specialized programs can extract from LKB only those pieces of information that are necessary, being presented in optimized formats, according to the requirements of the specific studies and/or applications. We list only a few applications: speller, grammar checker, inflection, synthetic and/or analytical morphological analysis, synthetic and/or analytical dependency tree representation, thesaurus, phonetic transcription, syllabification, lemmatization, diacritical checking/correction, document indexing and searching etc.

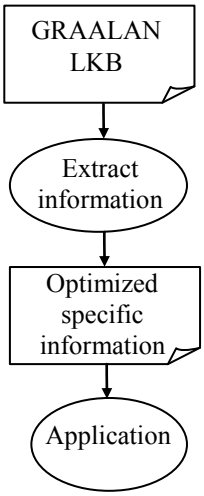

Fig. 3. LKB operations for a GRAALAN-based application

B. *The syllabification rules*

The syllabification rules are described by the linguist, using the GRAALAN metalanguage and starting from the information previously introduced in the section Alphabet (stress, letters, phonetic characters, diphthongs, phonetic groups, special characters, morphemes etc.) [17]. The GRAALAN system allows the description of three types of syllabification – euphonic, phonetic, and morphologic – with the following characteristics:

- *Euphonic syllabification* – phonetic syllabification rules applied to the alphabetical (normal) form of the word
- *Phonetic syllabification* – phonetic syllabification rules applied to the phonetic form of the word [18]
- *Morphologic syllabification* – morphologic syllabification rules applied to the alphabetical (normal) form of the word

For example, the lemma *dreptunghi* (rectangle) has associated the following syllabifications: *drep/tunghi* (euphonic), *drep.t'uŋ* (phonetic), *drept-unghi* (morphologic).

C. *The inflection process*

In order to automatically obtain the paradigm of a lemma (i.e. the set of all its inflected forms), the linguist describes a set of inflection rules, using the GRAALAN metalanguage. Thus, the collection of inflection rules for a certain language consists of all the transformations applied to the basic form of a word (the canonical form or lemma) in order to obtain all the inflection situations corresponding to that word [19] [20]. For instance, the Romanian language is a highly inflected language [21] [22] [23].

An inflection situation is represented as an inflection tree, having as its root a morphological class (noun, verb, pronoun, etc.) and as its leaves morphological categories (number, tense, mood etc.) and values (singular, plural, definite, perfect etc.). Each inflection form has associated an inflection situation, represented as a label. For example, the form *mame* (mothers) has the inflection situation *SubstComAnimFemPlNomNeart* 'NCountIndComPlFemCom'. The collection of all the morphological categories and their corresponding values is represented as an Attribute Value Tree, called the Morphological Configurator.

An inflection rule describes the types of modifications applied to a lemma (and to all members of its morphological class) and consists of operations such as *delete*, *insert,* and *replace*. The inflection forms that are exceptions to the described rule will have associated a special inflection rule. In order to verify how the inflection rules have applied to the lemmas and to find the exceptions, the linguist uses an instrument, called MKT (Morphological Knowledge Tool). This tool generates all the inflection forms, based on the Morphological Configurator and allows the linguist to operate the necessary modifications and to create new inflection rules.

An inflection rule describes the transformations to be applied on a lemma both on alphabetic and phonetic forms. Therefore, the paradigm of a word consists of all its inflection forms, written in normal and phonetic alphabet.

IV. GENERATING THE LIST OF SYLLABLES AND THE LISTS OF SYLLABLE SITUATIONS FROM A GRAALAN DATABASE

In order to create a representative syllable base for Romanian language, we start from single-word inflected forms. The GRAALAN linguistic database contains almost 800 000 single-word inflected forms, obtained by applying the inflection rules to almost 80 000 lemmas. Thus, the syllable base will cover a very-large portion of the Romanian words (contained in an explanatory dictionary, like DEX [14]), not only the most frequent words found in a corpus. The grammatical and deterministic approach ensures a minimal error rate of the future speech recognition system.

The list of inflected forms is made of the following data for every inflected form:

- the inflected form in normal alphabet
- the phonetic transcription (IPA - International Phonetic Alphabet) of the inflected form
- *the syllabification of the inflected form in normal alphabet*
- *the syllabification of the inflected form in phonetic alphabet*

In the process flow of a Syllable-based Speech Recognition System, a syllable is first identified by its acoustic features and thus by its phonetic alphabet transcription. Therefore an audio segment recognized as a potential phonetic alphabet transcription of a syllable can have one or more interpretations in normal alphabet transcription. In other words, a phonetic alphabet transcription corresponds

to one or more normal alphabet transcriptions. Similarly, the inverse multiple connection also exists.

The list of inflected forms can provide various other data necessary to the syllable analysis (see Fig. 4):

- the list of the distinct syllables of the analyzed language, extracted in normal and phonetic alphabet (SL – Syllable List)
- a list of distinct words that cover all the syllables of the analyzed language (SWL – Syllable Word List), extracted in normal and phonetic alphabet
- the list of distinct syllable adjacencies (AL – Adjacency List) of the analyzed language, extracted in normal and phonetic alphabet
- a list of distinct words, that cover all the syllable adjacencies (AWL – Adjacency Word List) of the analyzed language, extracted in normal and phonetic alphabet

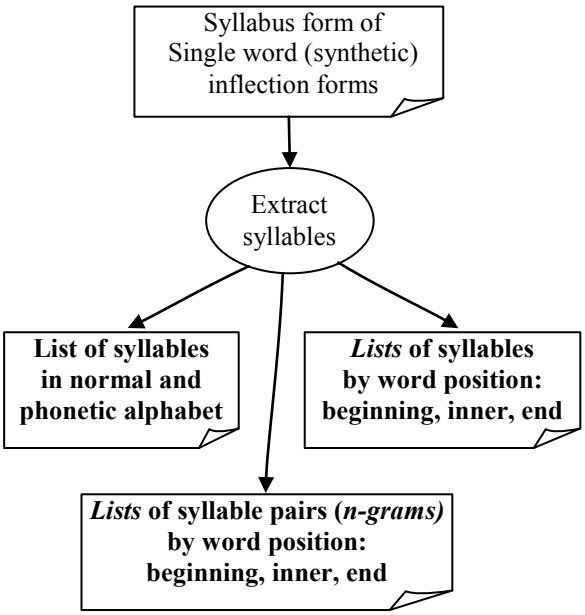

Fig. 4. Generate the list of syllables and the lists of syllable situations

A. A syllable can be in one of the following relations to the word containing it:

- isolated (monosyllabic word)
- left (the first syllable in a word with more than one syllable)
- inner (not the first, not the last syllable in a word with at least three syllables)
- right (the last syllable in a word with more than one syllable)

B. In relation to other words, a syllable can be in the following situations:

1) The case of an isolated syllable (monosyllabic word) – see A

- Preceded by the last syllable of another word and followed by the first syllable of another word (in case of a monosyllabic word)
- Preceded by the last syllable of another word and followed by silence (if isolated syllable)
- Preceded by silence and followed by the first syllable of another word (if isolated syllable)
- Preceded by silence and followed by silence (if isolated syllable)

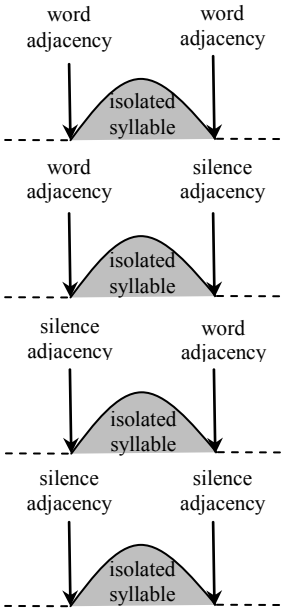

Fig. 5. Position and adjacencies of a syllable: a) The case of an isolated syllable (monosyllabic word)

2) The case of an inner syllable (minimum trisyllabic word) – see Fig. 6

- Preceded by a syllable of the same word and followed the a syllable of the same word

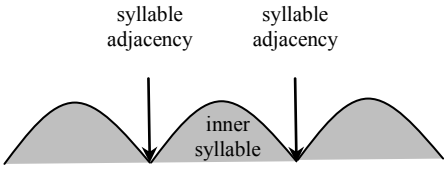

Fig. 6. Position and adjacencies of a syllable: b) The case of an inner syllable (minimum trisyllabic word)

3) The case of a left syllable (minimum bisyllabic word) – see Fig. 7

- Preceded by the last syllable of a word and followed by a syllable of the same word
- Preceded by silence and followed by a syllable of the same word

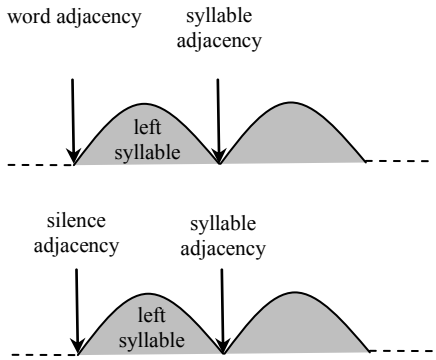

Fig. 7. Position and adjacencies of a syllable: c) The case of a left syllable (minimum bisyllabic word)

4) The case of a right syllable (minimum bisyllabic word) – see Fig. 8

- Preceded by the penultimate syllable of the same word and followed by the first syllable of another word
- Preceded by the penultimate syllable of the same word and followed by silence

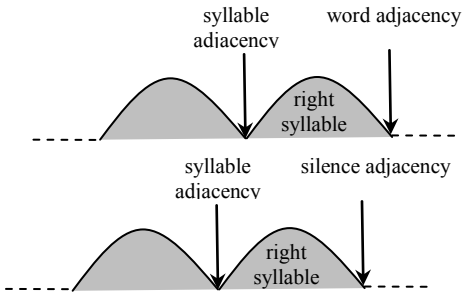

Fig. 8. Position and adjacencies of a syllable: d) The case of a right syllable (minimum bisyllabic word)

The audio base of syllable is designed to contain recordings that cover all the syllables found in the list of single-word inflected forms, but also recordings that cover different positions in which a syllable can occur in a word. Therefore, for every Romanian syllable we will have an example of a word for each of the three possible positions: the beginning, the inner, and the end of the word (as in Fig. 9).

a, a, (B) a/baţi, a/bˈatsʲ (I) bo/a/bab, bo/a/bˈab (E) bri/a, brˈi/a
a, ˈa, (B) a/bur, ˈa/bur (I) bo/a/rul, bo/ˈa/rul (E) cre/a, kre/ˈa
a, ˌa, (B) a/e/ro/lit, ˌa/e/ro/lˈit (I) his/to/a/u/to/ra/di/o/gra/fi/e, his/to/ˌa/u/to/ra/di/o/gra/fˈi/e
ab, ab, (B) ab/ces, ab/ʧˈes (I) ne/ab/stras, ne/ab/strˈas
ac, ak, (B) ac/tant, ak/tˈant (I) re/ac/tant, re/ak/tˈant (E) da/i/ac, da/ˈi/ak
ac, ˈak, (B) ac/tul, ˈak/tul (I) re/ac/ţii, re/ˈak/tsij (E) di/ac, di/ˈak
aci, ˈaʧ, (E) bos/ni/aci, bos/ni/ˈaʧ
aci, aʧ, (E) da/i/aci, da/ˈi/aʧ
acşi, ˈakʃʲ, (E) mo/no/acşi, mo/no/ˈakʃʲ
act, ˈakt, (B) act, ˈakt (E) act, ˈakt
ad, ˈad, (E) as/cle/pi/ad, as/kle/pi/ˈad
ad, ad, (B) ad/junct, ad/ʒˈunkt (I) ne/ad/mis, ne/ad/mˈis
af, af, (B) af/gan, af/gˈan
af, ˈaf, (B) af/tă, ˈaf/tə
ag, ˈag, (E) pa/li/ag, pa/li/ˈag
ag, ag, (B) ag/nat, ag/nˈat (I) di/ag/nos/ti/ci/an, di/ag/nos/ti/ʧi/ˈan
ˈadʒ, agi, (B) agi, ˈadʒ (E) agi, ˈadʒ
a, a, (B) a/baţi, a/bˈatsʲ (I) bo/a/bab, bo/a/bˈab (E) bri/a, brˈi/a
ˈa, a, (B) a/bur, ˈa/bur (I) bo/a/rul, bo/ˈa/rul (E) cre/a, kre/ˈa
ˌa, a, (B) a/e/ro/lit, ˌa/e/ro/lˈit (I) his/to/a/u/to/ra/di/o/gra/fi/e, his/to/ˌa/u/to/ra/di/o/gra/fˈi/e
aʃ, aş, (B) aş/tept, aʃ/tˈept (I) ne/aş/tep/tat, ne/aʃ/tep/tˈat
ˈaʃ, aş, (B) aş/tri, ˈaʃ/tri (I) pi/aş/tri, pi/ˈaʃ/tri (E) gu/aş, gu/ˈaʃ
aʃʲ, aşi, (E) pso/aşi, psˈo/aʃʲ
ˈaʃʲ, aşi, (B) aşi, ˈaʃʲ (E) cre/aşi, kre/ˈaʃʲ
ˈaʃtʲ, aşti, (E) ci/ne/aşti, ʧi/ne/ˈaʃtʲ
ˈaʧ, aci, (E) bos/ni/aci, bos/ni/ˈaʧ
aʧ, aci, (E) da/i/aci, da/ˈi/aʧ

Fig. 9. Examples of syllables and words containing them on different positions (B) - Beginning, (I) - Inner, (E) - End

V. GENERATING THE MINIMAL LIST OF WORDS FOR FULL COVERAGE OF SYLLABLE SITUATIONS

In order to form the preliminaries for an audio base of syllables, audio-recording will focus on lists of example words that cover most of the syllable situations identified in section IV. For this process (see Fig. 10), the syllables are differentiated only by their phonetical alphabet transcription. Therefore, we have the following lists of words:

- WT - a list of words that covers the list of syllables
- WA1 - a list of words that covers the list of each (possible) syllable on the first position
- WB1 - a list of words that covers the list of each (possible) syllable on an inner position
- WC1 - a list of words that covers the list of each (possible) syllable on the last position
- WD1 - a list of monosyllabic words
- WA2 - a list of words that covers the list of each (possible) syllable pair on the first position.
- WB2 - a list of words that covers the list of each (possible) syllable pair on an inner position.

- WC2 - a list of words that covers each (possible) syllable pair on the last position.
- WD2 - a list of bisyllabic words

The choosing of words in these lists can be made in a random fashion in order to simplify the generation process. However, criteria can be imagined in order to satisfy certain objectives or to improve the recognition rate, as in:

- include lists of most used words with different scope
- prioritize long words, thus many syllables per word and lower recording effort
- ensure a minimal number of example words for each syllable situation

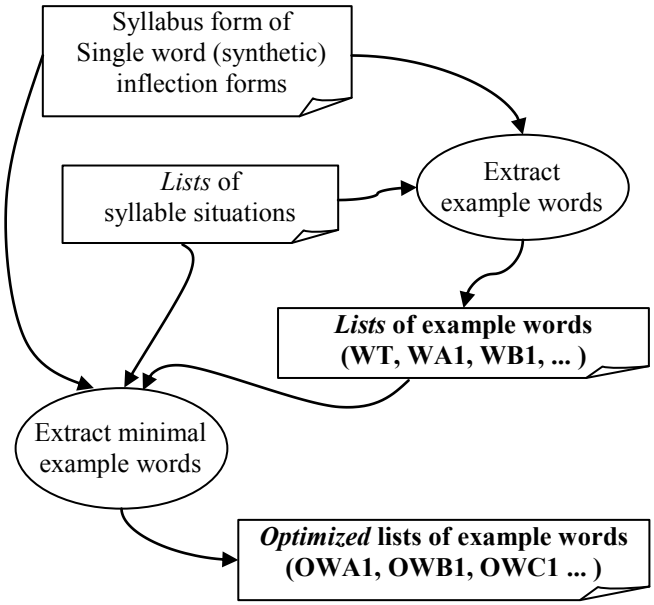

Fig. 10. Generate the lists of example words for syllable situations

These lists are not exclusive (some words can be found in several lists). From these lists we will get other optimized lists, i.e. lists from which the duplicates were removed: OWA1, OWB1, OWC1, OWD1, OWA2, OWB2, OWC2, OWD2. Optimizing the resulted lists of examples of words can be made taking into account the following criteria:

1. Generate a WT list of example words
2. Generate a WD1 list of example words with the WT list as filter for syllable situations
3. Generate a WA1 list of example words with the WT and WD1 lists as filter for syllable situations
4. Generate a WC1 list of example words with the previous lists (WT, WD1, WA1) as filter for syllable situations
5. Generate a WB1 list of example words with the previous lists (WT, WD1, WA1, WC1) as filter for syllable situations
6. Generate a WD2 list of example words with the previous lists (WT, WD1, WA1, WC1, WB1) as filter for syllable situations
7. Generate a WA2 list of example words with the previous lists (WT, WD1, WA1, WC1, WB1, WD2) as filter for syllable situations
8. Generate a WC2 list of example words with the previous lists (WT, WD1, WA1, WC1, WB1, WD2, WA2) as filter for syllable situations
9. Generate a WB2 list of example words with the previous lists (WT, WD1, WA1, WC1, WB1, WD2, WC2) as filter for syllable situations

The resulted lists of examples of words will be recorded by as many speakers as possible.

VI. BUILDING THE AUDIO BASE

As the final goal of our project is to develop a syllable-based speech recognition system, several dedicated speech corpora for both training and evaluation of such a system have to be collected.

A phone-based, state-of-the-art speech recognition system is usually trained and evaluated with speech corpora comprising naturally pronounced phrases (audio files of 5 to 30 seconds comprising several words in a given context). In the case of a syllable-based speech recognition (SSR) system such speech corpora are not enough. To properly analyze the speech signal corresponding to various syllables and also the variations of the speech signal when the syllables are found in a certain context, one needs audio files segmented at syllable-level.

In this context, our efforts were directed towards collecting:

- a speech corpus comprising isolated syllables,
- a speech corpus comprising isolated words, and
- a speech corpus comprising naturally pronounced sentences labelled at word and syllable level.

Speech corpus acquisition can be approached in two ways: by collecting already recorded audio files and labelling them with different granularity levels or by selecting texts (syllables, words or sentences) relevant for the corpus to be created and recording them. We chose to use the second method because we wanted to have a very strict control of the texts to be recorded. This fact is very important, because our goal was to create a speech corpus that comprises all the syllables in Romanian language (for training the SSR system).

A. The isolated syllables and isolated words corpora

The isolated syllables corpus and the isolated words corpus were collected using a speech recording application developed by the Speech and Dialogue research laboratory especially for the purpose of recording single words pronounced normally and syllable by syllable. The application displays whole words

on the screen and highlights the various syllables composing these words at every two seconds. The user reads the highlighted syllable and waits for a new syllable to be highlighted. In the end the whole word is highlighted and the user reads it and records it.

Using this application and following the above procedure, a list of 10.000 words was fully or partially recorded by the speakers in a group of 17 speakers (7 males and 10 females). The words were fully recorded by 9 speakers and partially recorded by another 8 speakers resulting in a corpus of 110.000 utterances comprising one word each.

The syllables were fully recorded by 5 speakers, resulting in a corpus of 185.390 utterances comprising one syllable each, with 37.078 utterances per speaker.

B. The sentences corpus labelled at word and syllable level

The naturally pronounced sentences corpus was collected using an online speech recording application developed by the same research laboratory[1]. The recording process was done remotely, each speaker using its own computer and recording environment. However, the application was designed to verify the signal-to-noise (SNR) ratio of every recording and reject the ones with an SNR lower than 30dB. The recording process is as follows: the application displays a sentence on the screen, the user presses a button to start the recording, the user reads the sentence out loud, the user presses another button to finish the recording, and finally the application sends the audio file to a server and displays the next phrase.

Using this procedure, 10 speakers partially recorded a list of 680 sentences selected from a Romanian novel ("Viața ca o pradă" written by Marin Preda). Each speaker recorded a sub-list of 123 to 144 sentences from the total of 680, each sub-list comprising 1671 to 1774 words. The corpus collected using this procedure comprises 1360 utterances with a total of 17292 words.

Following the collection procedure, the 1360 utterances were forced aligned using an automatic speech recognition system previously developed by the Speech and Dialogue research laboratory [24]. The goal of the forced alignment process was to automatically obtain timestamps for the words composing the utterances. These timestamps could be further used to split the utterances and obtain shorter utterances comprising a single word each. The benefit of this procedure over the one presented before for isolated words is that the output comprises utterances of words pronounced naturally in a sentence context.

Syllable-level timestamps for the naturally pronounced sentences corpus were also needed for the development and evaluation of the syllable-based speech recognition system. A first attempt to obtain these timestamps was done using the same speech recognition system mentioned above [24] and the same forced alignment method. This time the forced alignment was performed at phone-level producing phone-level timestamps. These timestamps were finally agglutinated to obtain syllable-level timestamps. However, a manual validation of the correctness of the timestamps showed that the results were not satisfactory (as in the case of word-level timestamps) and, consequently, the task was approached manually. The open-source WaveSurfer[2] application for sound visualization and manipulation was used for this task. The tool was used to display the waveform and the spectrogram of the speech signal. This visual information was used by the operator to place timestamps for each syllable in the sentences corpus. So far this is work in progress, as only half of the corpus is currently labeled at syllable level.

VII. STATISTICS FOR ROMANIAN LANGUAGE

As stated before, in our view, building a representative base of syllables starts from a linguistic database. This process is usually time consuming and implies a phasing approach. Combining this with an action to substantially extend the linguistic knowledge can result in a degree of error over the following reported statistics.

In TABLE I. we present some general statistics regarding the syllables for the Romanian linguistic database (from a GRAALAN viewpoint). These number represent the state of the linguistic database at a point in time before the large extension of the linguistic database in the AFLR project (*Romanian Language Phonetic Analysis: Study and applications*).

TABLE I. OVERVIEW OVER ROMANIAN SYLLABLES (GRAALAN PERSPECTIVE)

Entities	Number
Lemmas	**76 840**
Single-word inflected forms	**778 868**
Distinct syllables	**11 363**
Distinct adjacencies of syllables	**134 433**
Total number of syllables for all the single-word inflected forms	**3 446 945**
Maximum number of syllables in a word	**12**
Average number of syllables in a word	**4**

In the first phase of recording, 10 000 example words have been recorded for 10 000 distinct syllables, in alphabetic-phonetic pair. These recordings were made by 5 speakers, both using continuous speech and word syllabification. A number of 37 078 audio files resulted, each containing a syllable, a corpus that covers 10 909 unique syllables in alphabetic-phonetic pair. For example, the syllable with most recordings is '*le*', with a total of 1530 recordings. 9313 syllables are represented only by a single recording.

[1] Speech Recorder: http://speed.pub.ro/speech-recorder

[2] WaveSurfer: https://sourceforge.net/projects/wavesurfer

TABLE II. COUNTERS FOR SPECIFIC LISTS OF ROMANIAN SYLLABLES

List Name	List description	Dimensions (number of distinct situations)
WA1	The list of words for the possible situation in which every syllable occurs on the **start position** in a word	**3794** distinct syllables
WB1	The list of words for the possible situation in which every syllable occurs on an **inner position** in a word	**3391** distinct syllables
WC1	The list of words for the possible situation in which every syllable occurs on the **end position** in a word	**8180** distinct syllables
WD1	The list of words for the possible situation in which every syllable forms a **monosyllabic word**	**3014** distinct syllables
WA2	The list of words for the possible situation in which every syllable pair occurs on the **start position** in a word	**66158** distinct syllble pairs
WB2	The list of words for the possible situation in which every syllable pair occurs on an **inner position** in a word	**38034** distinct syllable pairs
WC2	The list of words for the possible situation in which every syllable pair occurs on the **end position** in a word	**88589** distinct syllable pairs
WD2	The list of words for the possible situation in which every syllable pair forms a **bisyllabic word**	**38489** distinct syllable pairs

For the 8 types of syllable situation described in section IV, the Romanian language has the counters found in TABLE II. .

VIII. CONCLUSIONS

This paper describes the work involved in building a *wide coverage audio base of syllables* for Romanian.

The presented work started from previous results, a wide coverage vocabulary and its syllabus form for Romanian, developed using the GRAALAN metalanguage and the corresponding linguistic knowledge bases management tools. This effort was part of the AFLR (*Romanian Language Phonetic Analysis: Study and applications*) [25] research project that aims to develop complex phonetic knowledge bases, including rules for automatic phonetic transcriptions and syllabifications.

Among the potential future exploitations of this research we could name: automatic synthetic syllabification of Romanian words (based on a set of general rules and algorithms that could also be applied for other languages), an inventory of the Romanian syllables, statistical studies on the frequency of the phonological units in Romanian texts, lexical and morphological studies about the migration of accents in the paradigm of a word, studies on the phonological sequence of words, and so on. Moreover, the complex audio base of syllables we described here will also represent an important resource for Automatic Speech Recognition.

Regarding the resource availability, some additional descriptions and a version of the phonetic and morphologic large vocabulary and its syllabus form for Romanian can be found in the previous published work [17][21][22][18][23]. As for the audio base of syllables for Romanian or other language resources for Romanian one can reach out the authors for details.

Acknowledgment

This work has been realized through the Partnerships Program in priority areas - PNII, developed with the support of MEN - UEFISCDI, project no. PN-II-PT-PCCA-2013-4-1451, contract no. 332/2014.

References

[1] C. Ungurean, D. Burileanu, "An advanced NLP framework for highquality Text-to-Speech synthesis," SpeD 2011, pp. 1-6, Braşov, Romania

[2] B. H. Juang and L. R. Rabiner, "Automatic speech recognition - A brief history of the technology development", in Elsevier Encyclopedia of Language and Linguistics, Amsterdam, The Netherlands: Elsevier, 2005

[3] X. Huang, J. Baker; R. Reddy, "A Historical Perspective of Speech Recognition", in Communications of the ACM, Volume 57 Issue 1, January 2014, pp.94-103, DOI: 10.1145/2500887

[4] G. Hinton et al, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", IEEE Signal Processing Magazine, 29(6):82–97, 2012.

[5] C. S. Petrea, A. Buzo, H. Cucu, M. Pasca and C. Burileanu, "Speech Recognition Experimental Results for Romanian Language," in Proceedings of ECIT, 2010.

[6] Izhak Shafran, Mari Ostendorf, "Acoustic model clustering based on syllable structure", article in Computer Speech & Language 17 (4) : 311-328 May 2001

[7] O. Buza, G. Toderean, J. Domokos "A Rule-Based Approach to Build a Text-to-Speech System for Romanian", the 8th International Conference on Communications COMM 2010, ISBN 978-1-4244-6363-3, June 10-12, 2010, pp.83-86

[8] G. Toderean, O. Buza, J. Domokos, "Achievements in the field of voice synthesis for Romanian", in the Proceedings of the 8th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2015; ISBN: 978-146737560-3, DOI: 10.1109/SPED.2015.7343078

[9] Z. Hu, J. Schalkwyk, E. Barnard and R.Cole, "Speech recognition using syllable-like units", in Spoken Language, 1996. ICSLP 96. Proceedins., 1996

[10] D. T. Ives, D. R. R. Smith, R. D. Patterson, "Discrimination of speaker size from syllable phrases." J. Acoust. Soc. Am., 118, p.3816-3822.

[11] Carolyn Richie, Sarah Warburton, Megan Carter, "Audiovisual Database of Spoken American English", LDC2009V01. Web Download. Philadelphia: Linguistic Data Consortium, 2009

[12] Ana Barbu, "Inflected and Syllabic Forms Dictionaries," in Proceedings of Language Resources and Evaluation Conference (LREC2008), ISBN: 2-9517408-4-0, on-line http://www.lrec-conf.org/proceedings/lrec2008/, Marrakech, Maroc, 26-31 May, 2008.

[13] David Crystal, "A Dictionary of Linguistics and Phonetics, sixth edition", Blackwell Publishing, 2008

[14] Institutul de Lingvistică Iordgu Iordan – Alexandru Rosseti al Academiei Române, “Dicţionarul explicativ al limbii române, second edition”, 1998, Univers Enciclopedic, Bucharest

[15] S. Diaconescu: “Complex Natural Language Processing System Architecture”, in Corneliu Burileanu, Horia-Nicolai Teodorescu (Eds.), Advances in Spoken Language Technology, The Publishing House of the Romanian Academy, Bucharest 2007, pp. 228-240

[16] S. Diaconescu: “GRAALAN – Grammar Abstract Language Basics”, in GESTS International Transaction on Computer Science and Engineering, Vol.10, No.1, 2005

[17] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, “Fonetica limbii române”, (Vol. I, Elementele metalimbajului GRAALAN), ISBN 978-1514314784, 2015

[18] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, “Fonetica Limbii Române”, (Vol. IV, Dicţionarul fonetic al silabelor limbii române şi dicţionarul fonetic de rime al limbii române), ISBN 978-1514315422, 2015

[19] Ş. S. Diaconescu, C. Ingineru, M. Mateescu, F. Codîrlaşu, M. M. Rizea, G. Masei, "Inflection Tools for Natural Languages" (US patent application 12/486597)

[20] Ş. S. Diaconescu, C. Ingineru, F. C. Codîrlasu, M. M. Rizea, O. Bulibaşa, “General System for Normal and Phonetic Inflection”, in Corneliu Burileanu, Horia-Nicolai Teodorescu (Eds.), From Speech Processing to Spoken Language Technology, The Publishing House of the Romanian Academy, Bucharest 2009, pp.149-160

[21] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, “Fonetica limbii române”, (Vol. II, Dicţionarul morfologic şi fonetic al limbii române, A-L), ISBN 978-1514315125, 2015

[22] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, “Fonetica Limbii Române”, (Vol. III, Dicţionarul morfologic şi fonetic al limbii române, M-Z), ISBN 978-1514315262, 2015

[23] Şt. S. Diaconescu, M. M. Rizea, F. C. Codîrlaşu, M. Ionescu, M. Rădulescu, A.Mincă, Şt. Fulea: “Methods for Automatic Generation of GRAALAN-based Phonetic Databases” - in the Proceedings of the 8th Conference on Speech Technology and Human-Computer Dialogue (SpeD2015), Bucharest, 2015, pp. 135-142, ISBN: 978-1-4673-7560-3

[24] H. Cucu, A. Buzo, L. Petrică, D. Burileanu and C. Burileanu, “Recent Improvements of the SpeeD Romanian LVCSR System“, in the Proceedings of the 10th International Conference on Communications (COMM), Bucharest, 2014, pp. 111-114

[25] http://www.softwinresearch.ro/index.php/en/research-projects/aflr