Published in 2015 International Conference (8th) on Speech Technology and Human-Computer Dialogue (SpeD)

Methods for Automatic Generation of GRAALAN-based Phonetic Databases

Ştefan - Stelian Diaconescu, Monica - Mihaela Rizea, Felicia - Carmen Codîrlaşu, Mihaela Ionescu,
Monica Rădulescu, Andrei Mincă, Ştefan Fulea
Research & Development Department
SOFTWIN
Bucharest, Romania

***Abstract*—This paper presents methods for automatic generation of phonetic databases (*The Morphological and Phonetic Dictionary, The Phonetic Dictionary of Syllables, The Rhyming Dictionary*) for a natural language, starting from a set of linguistic knowledge bases. The knowledge bases are developed by means of the GRAALAN (Grammar Abstract Language) system. The exemplification of this process will be described by representing a Romanian Phonetic database.**

***Keywords*—phonetics; linguistics; natural language processing**

I. Introduction

The automatic generation of large-scale phonetic databases (in the form of machine-readable and even human-readable dictionaries that record phonetic transcriptions and syllable divisions/stress patterns using an internationally-approved standard) is an important task, especially in the context of languages such as Romanian, which are considered under-resourced [1], [2]. The current state of research in the domain of Romanian phonetics and phonology underlines the necessity of a number of applied studies with the aim of obtaining (formalized) rules for automatic phonetic transcriptions and automatic syllabifications; the principles of this process, that is applicable to all natural languages, along with the complex structure of the GRAALAN (Grammar Abstract Language) system will be described in the following pages.

The rules for grapheme-phoneme correspondence are rendered in the form of GRAALAN-encoded predefined groups, by means of which a series of contexts, representative for the letter-sound pairing characteristic to each language, are identified. In the process of phonetic transcription, the linguistic teams select the predefined GRAALAN groups by making use of internally developed applications. All groups that contain vocalic sounds also include representations of primary and secondary stress. The phonetic dimension of the GRAALAN groups (as well as the entire phonetic database resulted) is represented in the International Phonetic Alphabet (IPA) standard, a very common representation of the phonetic alphabet in linguistic studies, which is equally computer-readable once the Unicode support for IPA symbols became widespread (for IPA in Unicode see [3]). The IPA transcription is rendered with the standard SIL font, widely used in phonetic word-processing all over the world [4]. The latest development of Unicode as well as our (semi-automatic) method of generating the phonetic database excluded the need of using other computer-readable systems for representing the IPA in ASCII (such as SAMPA and its variants). The resulting phonetic dictionaries generally make use of the principle of broad phonetic transcription (also named phonemic transcription), following the tradition of similar works of this kind (see, for example, Roach 2000 and the CEPD - Cambridge English Pronouncing Dictionary).

Syllable boundaries are also generated by means of GRAALAN-encoded sets of rules, specific for each natural language described, and in agreement with the generally approved linguistic standards (specified by authoritative linguistic studies). In the case of Romanian language, our reference paper is DOOM 2005, an important resource were the rules (and exceptions) for correct syllabification in Romanian are specified. In our GRAALAN treatment, we dealt with exceptions by means of the algorithm implemented, which ensures that the most specific variant of a syllabification rule is applied last. Romanian represents two types of syllabification (understood here as **results** of the syllabification process) in GRAALAN: Phonetic and Euphonic syllabification. Phonetic syllabification **rules** (based on pronunciation principles) are applied to words transcribed in the IPA standard. The Euphonic syllabification (which applies to words in their alphabetic form) is generated only afterwards. This mechanism was especially convenient for Romanian where the current norm recommends the pronunciation principles, i.e. the pronunciation principles take precedence over the structural (morphological) principles (DOOM 2005: LXXX). The situation is completely different in a language such as English, where the morphological syllabification boundaries applied to the written form of the words are equally important as the phonetic syllabification boundaries that follow phonetic principles and are applied on the phonetic variants of the words.

The state of the art in the field of phonetic transcription implies a series of methods and approaches ranging from context-sensitive grammars [17], to Expectation-Maximization (EM) algorithms [18] and uses of Hidden Markov Models (HMMs) [19] [20]. One of the problems reported relates to the low degree of accuracy obtained in the case of languages with alphabetic (or non-phonemic) orthography, such as English

ISBN : 978-1-4673-7560-3
DOI : 10.1109/SPED.2015.7343082

(only 65% to 71% for English phonetic transcription of OOV (out of vocabulary) words [19]).

The semiautomatic method applied by making use of GRAALAN predefined groups and internally developed applications ensures a high degree of accuracy since the databases are filled exclusively by linguist experts.

Similar efforts for automatic generation of phonetic databases for Romanian language imply either rule-based systems [5] [6], machine learning systems based on artificial neural networks [7] [8], or hybrid systems (that use both transcription rules and machine learning in order to solve ambiguity issues) [9] [10] [11] [12].

It is important to mention here that recent approaches used for automatically creating phonetic transcriptions also apply statistical machine translation (SMT) principles. See, for example, [13] where graphemes are regarded as “words” (and sequences of graphemes as “phrases”) in the source language, while (sequences of) phonemes are treated as the “words”/ “phrases” of the target language. As the authors point out, a further improvement can be obtained by combining the statistical approach with the rule-based one.

Notable Romanian language Phonetic databases are “Database of the Romanian Language Phonetics and Phonology” [14], Text-to-Speech Synthesis for Romanian Language [15] (with more than with 4,000 syllables digitally stored), and NaviRO [16] containing a phonetic transcription database manually generated by linguistic experts (2,383 words) and also a phonetic transcription database automatically generated using a system made of 30 artificial neural networks containing over 138,500 words.

The phonetic databases reported for Romanian generally contain a reduced number of entries (a notable exception being the corpus of 11,819 manually-corrected syllabified words transformed in their phonetic forms, mentioned in [12]).

The Romanian Morphological and Phonetic Dictionary that we have developed contains phonetic transcriptions for approximately *80,000* lemmas, a number of approximately *900,000* synthetic inflection forms, and a number of approximately *14,000,000* analytical inflection forms. In addition, the Romanian Phonetic Dictionary of Syllables that we have developed contains approximately *12,000* syllables (stress marker variants included).

The following sections will present a detailed description of the methods we applied for obtaining automatic phonetic databases, and we will also describe our strategy of obtaining phonetic databases for Romanian.

II. Brief description of the GRAALAN system

The phonetic database presented in this paper is built starting from the description of a natural language using the GRAALAN system. Therefore we offer here an overview of this system [36].

GRAALAN system is a set consisting of (see Fig. 1):

a) A set of internal *standards* concerning the creation and representation of the linguistic knowledge of a natural language (and the correspondences between two natural languages – but we will not address these issues here). These standards are grouped into a metalanguage called GRAALAN (Grammar Abstract Language) which enables the formal description of all linguistic sections corresponding to a natural language: the alphabet, the syllabification rules, the morphology, the inflection rules, the lexicon, the inflection forms, the syntax, the correspondences between two natural languages.

b) A set of tools (programs) that help linguistic teams generate the descriptions mentioned above.

c) A linguistic knowledge base that hosts the linguistic descriptions (usually in XML format – *Extended Markup Language*). Various linguistic applications can be developed based on this formalized linguistic knowledge, including the phonetic database this paper deals with.

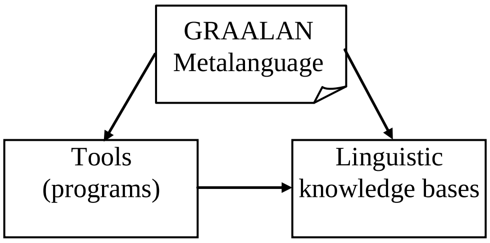

Fig. 1. Structure of GRAALAN system

An example of GRAALAN system operation comprises the following steps (see Fig. 2):

a) The linguist describes the following linguistic sections of a natural language, using GRAALAN: the alphabet, the syllabification rules, the morphology, the inflection rules, the lexicon, the syntax.

b) GRAALAN descriptions are then compiled by converting GRAALAN texts into XML format. At this point, the linguist can revise the descriptions based on the output generated by the GRAALAN compiler.

c) The linguist creates (mostly in RDB – *Relational Data Base*) the lexicon of the natural language using a special tool, called LKT (*Lexicon Knowledge Tool*).

d) The paradigms [the synthetic (single-word) and analytical (multi-word) inflection forms] of all lexicon lemmas are generated using the especially designed tool, called MKT (*Morphological Knowledge Tool*). MKT receives as input the alphabet, the morphological configurator and the inflection rules, all previously compiled and in XML format. The paradigms are generated by applying the inflection rules to the lexicon lemmas.

This is the process of creating an LKB (*Linguistic Knowledge Base).*

Subsequently, various applications can be developed based on the linguistic knowledge created. For this purpose, depending on the application requirements, only the information needed by the application will be extracted from the LKB (see Fig. 3). In addition, the structure and the representation of this information specific to the application will be optimized in order to better correspond (rapidly and with minimum resource consumption) to the application needs.

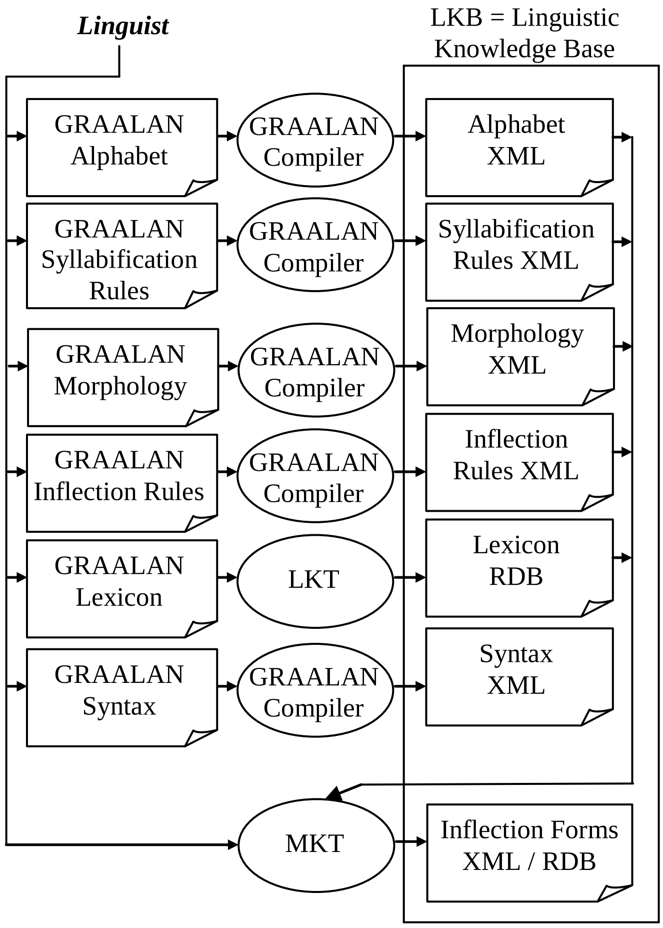

Fig. 2. Example of a GRAALAN system operation flow

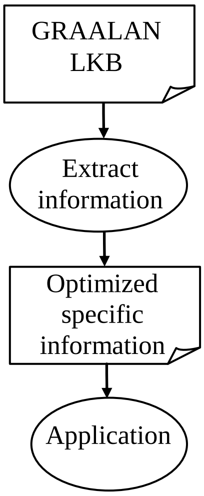

Fig. 3. LKB operations for a GRAALAN-based application

III. Tools for Building a GRAALAN Linguistic Knowledge Base

A set of tools (computer programs) were developed in order to help the linguist coping with the GRAALAN metalanguage, in order to facilitate his/her work and to handle a huge amount of linguistic knowledge. These tools include:

1) ***The GRAALAN Compiler***: a tool that converts the linguistic descriptions made by the linguist from GRAALAN (a format suitable for describing and understanding the linguistic information by the linguist) in XML format (*Extended Markup Language*) – a format suitable for representing and processing the information on the computer.

2) ***LKT (Lexicon Knowledge Tool)***: a tool that facilitates the development of the lexicon.

3) ***MKT (Morphologic Knowledge Tool)***: a tool that applies the inflection rules described by the linguist to the lexicon lemmas and thus obtains all the inflection forms (in normal and phonetic alphabet).

4) ***BDKT (Bilingual Dictionary Knowledge Tool)***: a tool that helps the linguist to create the correspondences between two natural languages.

5) ***LINK***: a tool that verifies the consistency of the linguistic knowledge bases.

IV. The Structure of the Phonetic Database of a Natural Language

We consider that the phonetic database of a natural language can consist of the following dictionaries (according to the current GRAALAN implementation):

A. The Morphological and Phonetic Dictionary

The Morphological and Phonetic Dictionary (see Fig. 4) should contain an entry for each word (synthetic or analytical form). The sorting can be done (or should be accessible if we are dealing with a computer application) according to the normal or the phonetic alphabet.

An entry in this dictionary includes (according to the current GRAALAN implementation):

- The word in normal alphabet.
- The word in phonetic alphabet.
- The word syllabification in normal alphabet.
- The word syllabification in phonetic alphabet.

- The morphological characterization of the word, as a sequence of morphological categories and morphological category values.

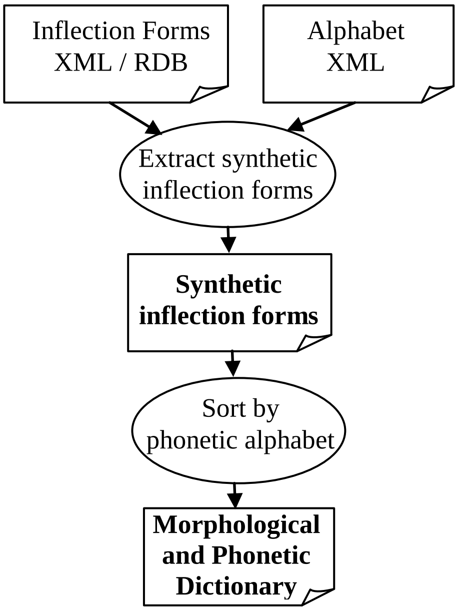
Inflection Forms XML / RDB
Alphabet XML
Extract synthetic inflection forms
Synthetic inflection forms
Sort by phonetic alphabet
Morphological and Phonetic Dictionary

Fig. 4. Building the *Morphological and Phonetic Dictionary*

B. The Phonetic Dictionary of Syllables

The *Phonetic Dictionary of Syllables* contains an entry for each syllable of the natural language. The sorting can be done according to the normal or the phonetic alphabet.

An entry in the *Dictionary of Syllables* includes (according to the current GRAALAN implementation):

- The normal form of the syllable
- The phonetic form of the syllable
- **(I)** An example of a word in which the syllable occurs at the beginning. If there is no word beginning with that syllable – nothing is written, not even **(I)**. If there are only monosyllabic words containing that syllable – **(I)** is written, followed by a monosyllabic word. If there are disyllabic (or with more syllables) words beginning with that syllable - **(I)** is written, followed by a disyllabic (or with more syllables) word.
- **(M)** An example of a word in which the syllables occurs in the middle. If there is no trisyllabic word containing that syllable – nothing is written, not even **(M)**. If there are minimum trisyllabic words containing that syllable - **(M)** is written, followed by a trisyllabic (or with more syllables) word.
- **(S)** An example of a word in which the syllables occurs at the end. If there is no word ending in that syllable – nothing is written, not even **(S)**. If there are only monosyllabic words containing that syllable - **(S)** is written, followed by a monosyllabic word. If there are disyllabic (or with more syllables) words ending in that syllable - **(S)** is written with a disyllabic (or with more syllables) word.

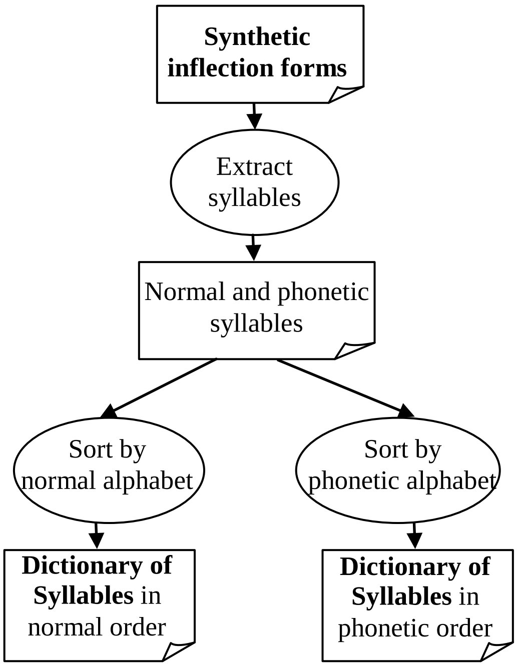
Synthetic inflection forms
Extract syllables
Normal and phonetic syllables
Sort by normal alphabet
Sort by phonetic alphabet
Dictionary of Syllables in normal order
Dictionary of Syllables in phonetic order

Fig. 5. Building the *Dictionary of Syllables*

The words given as examples are hyphenated, first written in normal alphabet and then in phonetic alphabet. In both cases, the syllables are separated by slash ("/").

C. The Rhyming Dictionary

A (simple) phonetic rhyme is defined as the string of phonemes from the last stressed vowel to the end of a word. (There are, of course, multiple rhymes and assonances, but we are not dealing with them here). The *Rhyming Dictionary* of a language contains all the words (synthetic and possibly analytical inflection forms) of that language, ordered by the rhyme, according to the phonetic alphabet.

An entry in this Rhyming Dictionary includes (according to the current GRAALAN implementation):

- The rhyme in phonetic alphabet.
- The rhyme in normal alphabet.
- The words containing that phonetic rhyme ordered and grouped by the number of syllables. Each group is preceded by the corresponding number of syllables written in brackets. Within each group, the words, (written in normal alphabet) are ordered according to the normal alphabet.

V. The generation of a natural language phonetic database

A. The Morphological and Phonetic Dictionary

The generation of the *Morphological and Phonetic Dictionary* (see Fig. 4) is done through the following steps:

1) Having as input two GRAALAN sections, the *Alphabet* (XML) and the *Inflection Forms* (XML), the process extracts all the synthetic inflection forms. Each entry has four text fields: the text (of the synthetic form) in normal alphabet, the text in phonetic alphabet, the text syllabified in normal alphabet, and the text syllabified in phonetic alphabet. Additionally, the morphological information will be specified for each entry (morphological categories and values of morphological categories).

2) The list of synthetic inflection forms will be sorted in phonetic order using the phonetic alphabet (over the field of phonetic text in each entry) and thus generating the *Morphological and Phonetic Dictionary*.

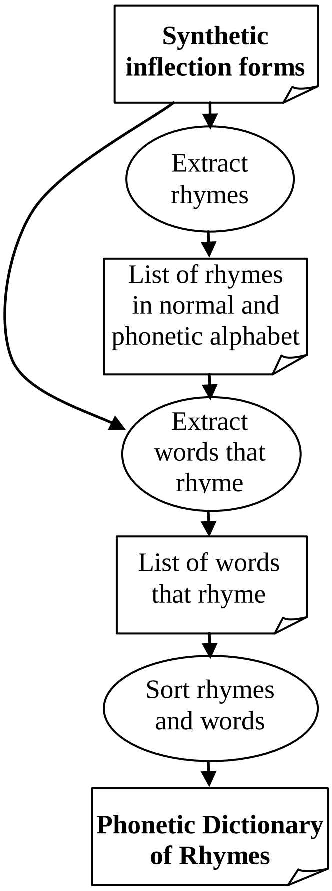
Synthetic inflection forms
Extract rhymes
List of rhymes in normal and phonetic alphabet
Extract words that rhyme
List of words that rhyme
Sort rhymes and words
Phonetic Dictionary of Rhymes

Fig. 6. Building the *Rhyming Dictionary*

B. The Phonetic Dictionary of Syllables

The generation of the *Phonetic Dictionary of Syllables* (see Fig. 5) is done through the following steps:

1) All syllables are extracted from the list of synthetic inflection forms and the list of distinct syllables is generated. These will also have two forms of textual representation: in normal and phonetic alphabet.

2) By sorting the syllables in accordance with the normal alphabet we obtain the *Dictionary of Syllables* in normal order.

3) By sorting the syllables in accordance with the phonetic alphabet we obtain the *Dictionary of Syllables* in phonetical order.

C. The Rhyming Dictionary

The generation of the *Rhyming Dictionary* (see Fig. 6) is done through the following steps:

1) Extract rhymes specific to all synthetic inflection forms. Each entry in this list will have one textual representation in phonetic alphabet and one or more textual representations in normal alphabet.

2) For each rhyme, extract the list of synthetic inflection forms that include this rhyme. The result is the list of synthetic inflection forms grouped by rhymes.

3) The resulted lists are ordered as follows:

a) first, the rhyme lists are sorted by the rhyme's text (the text in phonetic alphabet);

b) in each rhyme list the words are sorted and partitioned in subgroups by the number of syllables in each word;

c) in each rhyme subgroup the words are sorted by their normal textual representation (normal alphabet).

The *Phonetic Rhyming Dictionary* is thus obtained. Of course, in the case of specific applications other index and search mechanisms can be used in order to ease the access to words that rhyme.

VI. Exemplification of Romanian Language

A. The Romanian Morphological and Phonetic Dictionary

The *Romanian Morphological and Phonetic Dictionary* consists of approximately *900,000* distinct synthetic inflection forms and approximately *14,000,000* distinct analytical inflection forms, for a lexicon of approximately *80,000* lemmas (lexicon entries).

The Romanian GRAALAN-based morphological configurator, on which the above inflection situations are built, is made of *70* attributes and *257* attribute-values in a form of a tree structure (AVT - attribute value tree).

The GRAALAN-based Inflection Rules section for Romanian is formed of *260* basic rules and of *420* compound rules from which almost *280* compound rules are MKT-generated.

An example of an entry in such a dictionary is as follows [22][23]:

abecedar, abeʧedˈar, a/be/ce/dar, a/be/ʧe/dˈar [Cls.=Subst.] [TipSubst.=Com.] [Animat.=Inanim.] [Gen=Neu.] [Nr.=Sg.] [Caz=Nom.] [Art.=Neart.]

where:

a) ***abecedar*** = the word in normal alphabet.

b) ***abeʧedˈar*** = the word in phonetic alphabet (the stress marker is also present).

c) ***a/be/ce/dar*** = the word syllabified in normal alphabet.

d) ***a/be/ʧe/dˈar*** = the word syllabified in phonetic alphabet.

e) *[Cls.=Subst.] [TipSubst.=Com.] [Animat.=Inanim.] [Gen=Neu.] [Nr.=Sg.] [Caz=Nom.] [Art.=Neart.]* = morphological characterization of the word:

[Cls.=Subst.]	Class:	Noun
[TipSubst.=Com.]	Noun type:	Common
[Animat.=Inanim.]	Animation:	Inanimate
[Gen=Neu.]	Gender:	Neuter
[Nr.=Sg.]	Number:	Singular
[Caz=Nom.]	Case:	Nominative
[Art.=Neart.]	Article:	Non-articled

B. The Romanian Phonetic Dictionary of Syllables

The *Romanian Phonetic Dictionary of Syllables* contains approximately *12,000* syllables. Usually, a syllable occurs in two variants: the stressed form and the unstressed form. As a result, if we were to ignore the stressed / unstressed markers there will be only approximately *6,000* distinct syllables.

In what concerns the position of the syllables, for the Romanian language, the number of syllables found at the beginning of a word is approximately *3,800*. For the inner-word position there are approximately *3,400* syllables. Finally, at the end of the word position there are approximately *8,900* syllables.

An example of an entry in such a dictionary is as follows [24]:

blen, **blˈen**, **(I)** blen/de, blˈen/de **(M)** horn/blen/dă, horn/blˈen/də **(S)** go/blen, go/blˈen

where:

a) ***blen*** = The syllable in normal alphabet.

b) ***blˈen*** = The syllable in phonetic alphabet.

c) ***(I)*** = Precedes an example of a word in which the syllable takes the first position.

- *blen/de* = The syllable is placed in a word-initial position. The word is syllabified and in normal alphabet.
- *blˈen/de* = The syllable is placed in a word-initial position. The word is syllabified and in phonetic alphabet. The stress marker is present.

d) ***(M)*** = Precedes an example of a word in which the syllable takes an inner-word position.

- *horn/blen/dă* = The syllable is placed in an inner-word position. The word is syllabified and in normal alphabet.
- *horn/blˈen/də* = The syllable is placed in an inner-word position. The word is syllabified and in phonetic alphabet. The stress marker is present.

e) ***(S)*** = Precedes an example of a word in which the syllable take the end position.

- *go/blen* = The syllable is placed in a word-final position. The word is syllabified and in normal alphabet.
- *go/blˈen* = The syllable is placed in a word-final position. The word is syllabified and in phonetic alphabet. The stress marker is present.

C. The Romanian Rhyming Dictionary

In principle, the *Romanian Rhyming Dictionary* contains as many entries (i.e. synthetic inflection forms) as the number of rhymes. In the case of the analytical inflection forms, the number of words that rhyme can be considered larger.

In the latest GRAALAN-based LKB development [24] for the Romanian language, there are approximately *24,900* rhymes (as in group of words that rhyme).

An example of an entry in such a dictionary is as follows [24]:

ˈəʃtʲ (ăşti) (1) ăşti, băşti, căşti, găşti, hăşti, măşti, păşti, plăşti, tăşti, **(2)** adăşti, borăşti, bumăşti, chiorăşti, pălăşti, pârăşti, rubăşti, tălăşti, târăşti, urăşti **(3)** amărăşti, coptorăşti, cosorăşti, dogorăşti, hotărăşti, izvorăşti, mohorăşti, ocărăşti, ogorăşti, oţărăşti, oţerăşti, ponorăşti, scociorăşti, stoborăşti, tătărăşti, zădărăşti, zămorăşti, zăporăşti, zăvorăşti **(4)** deszăvorăşti, înzăvorăşti, posomorăşti

where:

a) ***ˈəʃtʲ*** = The rhyme written in phonetic alphabet *b)* ***(ăşti)*** = The rhyme written between brackets in normal alphabet (one or more normal representations).

c) ***(1)*** = Precedes the subgroup of words with one syllable.

- *ăşti, băşti, căşti, găşti, hăşti, măşti, păşti, plăşti, tăşti* = The words of the subgroup with just one syllable.

d) ***(2)*** = Precedes the subgroup of words with two syllables.

- *adăşti, borăşti, bumăşti, chiorăşti, pălăşti, pârăşti, rubăşti, tălăşti, târăşti, urăşti* = The words of the subgroup with two syllables.

e) ***(3)*** = Precedes the subgroup of words with three syllables.

- *amărăşti, coptorăşti, cosorăşti, dogorăşti, hotărăşti, izvorăşti, mohorăşti, ocărăşti, ogorăşti, oţerăşti, oţărăşti, ponorăşti, scociorăşti, stoborăşti, tătărăşti, zădărăşti, zămorăşti, zăporăşti, zăvorăşti* = The words of the subgroup with three syllables.

f) ***(4)*** = Precedes the subgroup of words with four syllables.

- *deszăvorăşti, înzăvorăşti, posomorăşti* = The words of the subgroup with four syllables.

VII. Conclusions

The system described in this paper is generally applicable to natural languages, and the results we are aiming at include the development of extended phonetic knowledge bases, large phonetic and morphological dictionaries as well as tools necessary for their completion. We also intend to develop applications of Automatic Speech Recognition that are based on acoustic elements (and make use of patented invariant-based algorithms) as well as on linguistic knowledge (phonetic transcriptions, syllabifications, spelling and grammar checking).

Equally important is the fact that, due to our methods applied for filling the phonetic databases, which exclusively imply annotation performed by linguistic experts, and to our strategies of verification at the end of each development stage, we can guarantee a high level of accuracy, which is considered a difficult task by other similar existing approaches (for example [39]).

Developing phonetic databases is a very important task for the evolution of spoken language technologies (SLT), including Automatic Speech Recognition (ASR) and Text-to-Speech (TTS), especially in the case of resource-scarce languages such as Romanian.

This phonetic database can also serve as an important teaching/learning instrument in the world of Romanian as a Foreign Language, as well as an important resource for comparative phonetic and phonological studies.

Acknowledgment

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

References

[1] C. Burileanu, V. Popescu, A. Buzo, C.S. Petrea and D. Ghelmez-Haneş, “Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems”, Proceedings of the Romanian Academy, Vol. 11, Series A, Number 1/2010, The Romanian Academy, Bucharest, Romania, pp. 83–91, 2010.

[2] D. Cristea and C. Forăscu, “Linguistic Resources and Technologies for Romanian Language”, Computer Science Journal of Moldova, vol.14, no.1 (40), Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova, pp. 34-73, 2006.

[3] https://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm

[4] http://scripts.sil.org/cms/scripts/page.php?item_id=IPAhome

[5] Ș.-A Toma, D. Munteanu, “Rule-Based Automatic Phonetic Transcription for the Romanian Language”, Proc. of the Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, Athens, pp. 682-686, 2009.

[6] T. Boroș, D. Ștefănescu, R. Ion, “Bermuda, a data-driven tool for phonetic transcription of words”, Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop, 35-39 (2012)

[7] D. Burileanu, “Basic Research and Implementation Decisions for a Text-to-Speech Synthesis System in Romanian”, International Journal of Speech Technology, Vol. 5, Springer, pp. 211-225, 2002.

[8] J. Domokos, O. Buza, G. Toderean, “Automated Grapheme-to-Phoneme Conversion System for Romanian”, Proceedings of the 6th Speech Technology and Human-Computer Dialogue Conference SpeD, 105-110 (2011)

[9] M. A. Ordean, A. Saupe, M. Ordean., M. Duma, G.C. Silaghi, “Enhanced Rule-Based Phonetic Transcription for the Romanian Language”, Proceedings of the 11th International Symphosium On Symbolic and Numeric Algorithms for Scientific Computation (SYNASC), 401-406 (2009)

[10] J. Domokos, O. Buza, G. Toderean, “100K+ words, machine-readable, pronunciation dictionary for the Romanian language”, Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 320-324 (2012)

[11] D. Jitcă, H.N. Teodorescu, V. Apopei, F. Grigoraş, “An ANN-based method to improve the phonetic transcription and prosody modules of a TTS system for the Romanian language,” SpeD 2003, pp. 43-50, Bucharest, Romania.

[12] C. Ungurean, D. Burileanu, “An advanced NLP framework for high-quality Text-to-Speech synthesis,” SpeD 2011, pp. 1-6, Braşov, Romania.

[13] H. Cucu, A. Buzo, L. Besacier, C. Burileanu, “SMT-based ASR Domain Adaptation Methods for Under-Resourced Languages: Application to Romanian“, in Speech Communication Journal, Vol. 56 – Special Issue on Processing Under-Resourced Languages, pp. 195-212, 2014.

[14] http://www.racai.ro/external/static/awde/serban.html

[15] http://www.racai.ro/external/static/awde/burileanu8.html

[16] http://users.utcluj.ro/~jdomokos/naviro/

[17] M. Divay and A. J. Vitale, “Algorithms for grapheme-phoneme translation for English and French: applications for database searches and speech synthesis”. In Computational Linguistics, 23(4), 1997, pp. 495–524.

[18] A. P. Dempster, N. M. Laird, D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”. In Journal of the Royal Statistical Society: Series B, 39(1), 1977, pp. 1–38.

[19] S. Jiampojamarn, C. Cherry, G. Kondrak, “Joint processing and discriminative training for letter-to-phoneme conversion”. In Proceedings of ACL-2008: Human Language Technology Conference, 2008, pp. 905–913, Columbus, Ohio.

[20] V. Demberg, “Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion”. In Proceedings of ACL-2007.

[21] 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, in press

[22] 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, in press

[23] 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, in press

[24] 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, in press

[25] Ş. S. Diaconescu, “Natural Language Understanding Using Generative Dependency Grammar”, in Max Bramer, Alun Preece and Frans Coenen (Eds), in Proceedings of SE2002, the 21-nd SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge UK, Springer, 439-452.

[26] 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

[27] Ş. S. Diaconescu, “Natural Language Syntax Description using Generative Dependency Grammar”, POLIBITS, Number 38, July-December 2008, ISSN: 1870-9044, pp. 5-18

[28] E. Vasiliu, “Scrierea limbii române în raport cu fonetica şi cu fonologia”, Bucureşti: Universitatea din Bucureşti, 1979

[29] A. Avram, “Semivocalele româneşti din punct de vedere fonologic”, Studii şi cercetări lingvistice 9, 1958.

[30] A. Graur and A. Rosetti, “Esquisse d’une phonologie du roumain”, Buletin Linguistique 6, 1938

[31] E. Petrovici, “Sistemul fonematic al limbii române”, Studii şi cercetări lingvistice 7, 1956

[32] I. Chiţoran, “The Phonology of Romanian: A Constraint-Based Approach”, Berlin; New York, Mouton de Gruyter, 2002

[33] I. Maddieson, “Phonetic Cues to Syllabification” in Phonetic linguistics: Essays in honor of Peter Ladefoged, Orlando, Florida, Academic Press, Inc., 1985

[34] D. Jones, P. Roach, J. Hartman, J. Setter and (editors), CEPD 2003, “Cambridge English Pronouncing Dictionary” CD-ROM, Cambridge: Cambridge University Press, 2003

[35] E. Vasiliu, A. Avram, “Les problèmes du système phonologique du roumain”, 1959

[36] Ş. S. Diaconescu, I. Dumitraşcu, C. Ingineru, O. Bulibaşa, M. M. Rizea, B. Păun, " System and methods for Natural Language Processing Including Morphological Analysis, Lemmatizing, Spell Checking, and Grammar Checking" (US patent, 8,762,130 B1 from 06/24/2014)

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

[38] Ş. S. Diaconescu, M. Mateescu, A.Mincă, G. Masei, B. Păun, "System for Managing a Complex Lexicon Comprising Multiword Expressions and Multiword Expression Templates" (US patent 8,762,131 B1 from 06/24/2014)

[39] D. József, O. Buza, and G. Toderean. "Romanian phonetic transcription dictionary for speeding up language technology development." Language Resources and Evaluation 49, no. 2 (2015): 311-325.