100K+ WORDS, MACHINE-READABLE, PRONUNCIATION DICTIONARY FOR THE ROMANIAN LANGUAGE

József Domokos, Ovidiu Buza, Gavril Toderean

Technical University of Cluj-Napoca, Communications Department

ABSTRACT

This paper intends to present a newly developed Romanian language pronunciation dictionary called NaviRo. The dictionary contains more than 100k words from the DexOnline dictionary together with their phonetic transcriptions in Speech Assessment Method Phonetic Alphabet (SAMPA), a machine readable alphabet. The development of the pronunciation dictionary and the system architecture are also described in the paper.
NaviRo pronunciation dictionary is freely available on the project website in HTK (Hidden Markov Model Toolkit) and Festival Speech Synthesis System dictionary format. There are also available for download the used grapheme and phoneme set and the audio samples for the used phonemes. The use of these resources is completely unrestricted for any research purposes in order to promote Romanian language speech technology research.

Index Terms— Romanian language speech recognition, speech synthesis pronunciation dictionary, grapheme-to-phoneme conversion, letter-to-sound conversion, phonetic transcription

1. INTRODUCTION

Pronunciation dictionaries are very useful resources for spoken language technology. These resourced are widely used in automatic speech recognition (ASR) and text to speech (TTS) synthesis applications [1] [2] because they are at the base of automated segmentation of speech at phonetic level [3], and also predicting the pronunciation of a written word is an important sub-task of all speech production systems [4]. In case of some languages such as Romanian, considered under-resourced [1][6], the existence of a pronunciation dictionary can considerably speed up ASR and TTS system development.

To our best knowledge there is not available any large, machine-readable pronunciation dictionary for Romanian language, as it is for example the Carnegie Mellon University (CMU) Pronouncing Dictionary [7] or the British English Example Pronunciations (BEEP) [8] for English, which can be used for grapheme-to-phoneme transcription in large vocabulary continuous speech recognition and text-to-speech system development.

The documentation studied [9]-[15] shows that there exist some automatic grapheme-to-phoneme transcription systems for Romanian, and also some small, hand built phonetically transcribed databases are reported which could be used for training and testing such systems, but these applications and resources are not freely available.

In the most important papers, grapheme-to-phoneme transcription for the Romanian language is handled using rule-based systems [12], neural network based machine learning systems [9]-[11][15] and hybrid systems that use transcription rules and machine learning to solve the ambiguities of rules [11][13].

This work is related to the NaviRo (http://users.utcluj.ro/~jdomokos/naviro) research project with the main objective to create a Romanian language voice driven navigation extension to the most popular WEB browsers like Mozilla Firefox, Opera and Google Chrome. An important subtask of this project is the development of a Romanian language pronunciation dictionary to be used for the recorded speech database phonetic transcription.

The scope of this paper is to present the newly developed NaviRO 100k+ words, machine-readable pronunciation dictionary for the Romanian language. The dictionary was developed in two stages. First using an artificial neuron network (ANN) system having a parallel structure of 30 neural networks (see **Fig. 2**) we have developed a 5k word pronunciation dictionary. The automated grapheme-to-phoneme transcription system was tested on a small 1k word, hand built database and previously presented in [15]. The trained system was able to perform grapheme-to-phoneme transcription with an accuracy of 92.83%, calculated at the phoneme level.

We then built the 100k+ words NaviRO dictionary using Dictionary Maker [16], a software application created to facilitate the creation of an electronic pronunciation dictionary in a target language starting from the previously created 5k dictionary and a 100k + wordlist.

The more than 100.000 Romanian words were collected from the DexOnline dictionary, the largest online dictionary for Romanian language [17]. DexOnline dictionary is freely available, can be downloaded from the

Internet and used in accordance with the terms of GNU General Public License.

The paper is organized as follows. Section 2 presents the used grapheme and phoneme set. Section 3 and 4 give details about the pronunciation dictionary development and testing. Section 5 summarizes our work and suggests some future works.

2. THE USED GRAPHEME AND PHONEME SET

The used grapheme set contains the 31 characters used for modern Romanian writing according to the second edition of DOOM – the spelling, orthoepic and morphological dictionary of Romanian language - [18] and is presented in *Table 1.*

Table 1. The 31 graphemes used for modern Romanian writing (according to [18])

a	**ă**	**â**	**b**	**c**	**d**
e	**f**	**g**	**h**	**i**	**î**
j	**k**	**l**	**m**	**n**	**o**
p	**q**	**r**	**s**	**ş**	**t**
ţ	**u**	**v**	**w**	**x**	**y**
z					

The Romanian phonetic inventory generally consists of 7 vowels, 2-4 semivowels and 20 consonants. *Table 2* shows the phone set used in our dictionary in SAMPA coding. It contains 32 phonemes including the phonetically null unit [sil].

Table 2. The used phoneme set presented in SAMPA coding

a	**@**	**1**	**b**	**k**	**d**
e	**e_X**	**f**	**g**	**h**	**i**
i_0	**j**	**l**	**m**	**n**	**o**
o_X	**p**	**r**	**s**	**S**	**t**
ts	**tS**	**u**	**v**	**z**	**Z**
dZ	**sil**				

In *Table 2* we have used the fallowing notations:

- **@** stands for Romanian **ă** grapheme (i.e. arm**ă**);
- **1** stands for Romanian **â** and **î** graphemes (i.e. rom**â**n, **î**nalt);
- **k** replaces character **c** (i.e. cot);
- **i_0** represents short **i** from the end of the words (i.e. lup**i**);
- **j** represents the grapheme **i** when pronounced as semivowel (i.e. **i**ar, oa**i**e);
- **S** is for Romanian character **ş** (i.e. şal);
- **ts** replaces character **ţ** (i.e. aţă);
- **tS** stands for the grapheme groups **ce**, **ci** (i.e. **ce**ramică, **ci**ocănitoare);
- **e_X**, **o_X** replaces the semivowels **ea** and **oa** (i.e. d**ea**l, s**ea**ră, s**oa**re);
- **dZ** for the grapheme groups **ge**, **gi** (**ge**m, **gi**n);
- **Z** stands for **j (**i.e. joi).

3. THE PRONUNCIATION DICTIONARY DEVELOPMENT

To develop the pronunciation dictionary we have used the largest online dictionary for Romanian language, the DexOnline – Online Explanatory Dictionary [17]. DexOnline dictionary is freely available and can be used in accordance with the terms of GNU General Public License. It can be downloaded from the Internet also as a mysqldump generated SQL file. The database can be easily restored on a MySQL relational database server. DEXOnline database is organized in multiple tables. The most important 3 tables from the point of view of exporting dictionary words are: *inflectedform*, *definition* and *lexem*.

The *inflectedform* table contains all the inflected forms of the words recorded in the database. By selecting al the distinct wordforms from this table we get a total number of 992.979 records. This is the maximum size of pronunciation dictionary we can create based on DEXOnline. We have exported these words in distinct text files separated by the first grapheme of the words, one word per line, thus resulting input files with a reasonable number of records in the order of several tens of thousands per file.

The *definition* table is a smaller table containing the definitions recorded in database. A total number of 126.563 definitions are available.

The *lexem* table contains just the base forms of words from the dictionary, totally 139.509. Some of these words are foreign language words and therefore they are not included in the pronunciation dictionary.

We have exported also these 2 tables in text format with UTF-8 character encoding, one word per line in order to perform automatic grapheme-to-phoneme conversion.

For the development of the 100k words pronunciation dictionary we have used the Dictionary Maker application [16], a software application created by the Human Language Technologies Research Group of the Meraka Institute, South Africa to facilitate the creation of an electronic pronunciation dictionary in a target language. A Dictionary Maker project can be started with either a word list or an initial dictionary or both. The word list defines the words that will be used to create the dictionary. For our experiments the word list consists of words exported from the DexOnline dictionary. Importing an initial dictionary will provide rules that can be used to predict pronunciations

for words in the word list. If the initial dictionary is small, the predictions will not be very accurate, but will improve through the bootstrapping verification process. If no dictionary is imported, there are no initial grapheme-to-phoneme prediction rules available, and therefore no initial pronunciation predictions will be given by the system. The user then has to provide the entire pronunciation. In this later case the development of a large pronunciation dictionary is a very time consuming task.

Dictionary Maker use a modified implementation of Kohonen's Dynamically Expanding Context (DEC) algorithm, a popular instance-based learning algorithm that predicts phoneme realization based only on grapheme context. DEC is used to extract rules of the form [5]:

(left context, grapheme, right context) ! phoneme.

Generating a new pronunciation is a simple procedure: each grapheme in the word is considered in turn, and the rule describing the largest matching context is used to predict the phoneme to be generated [5].

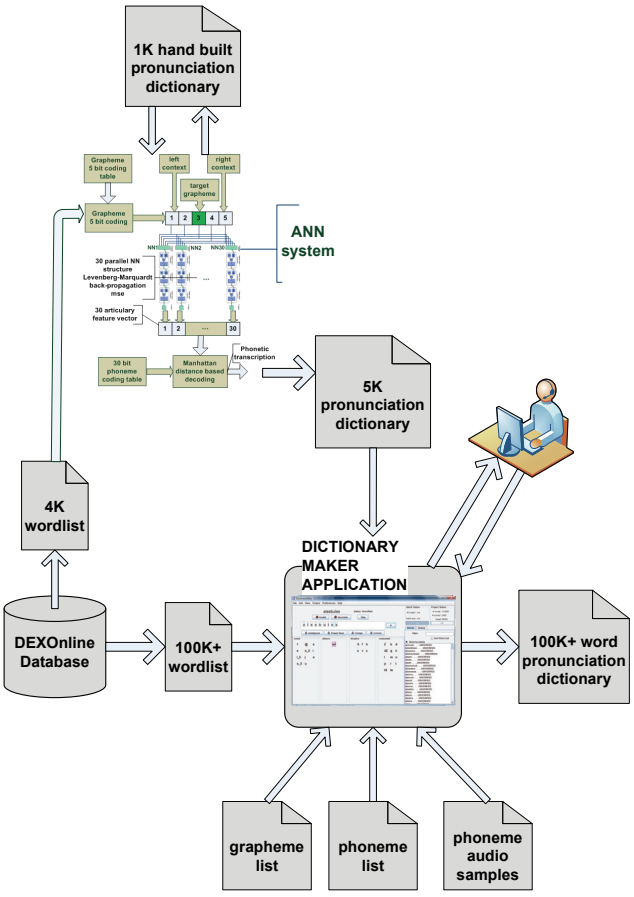

Fig. 1. System architecture for the development of the 100k + words pronunciation dictionary

We used Dictionary Maker with 5k transcribed words as initial dictionary. The 5k words were transcribed using our previously presented ANN based automated grapheme-to-phoneme transcription system [15]. In this way we have enough grapheme-to-phoneme prediction rules so that the transcriptions returned by Dictionary Maker do not need many corrections. Changes are needed only for exceptions from the standard pronunciation. The system architecture is depicted in **Fig. 1**.

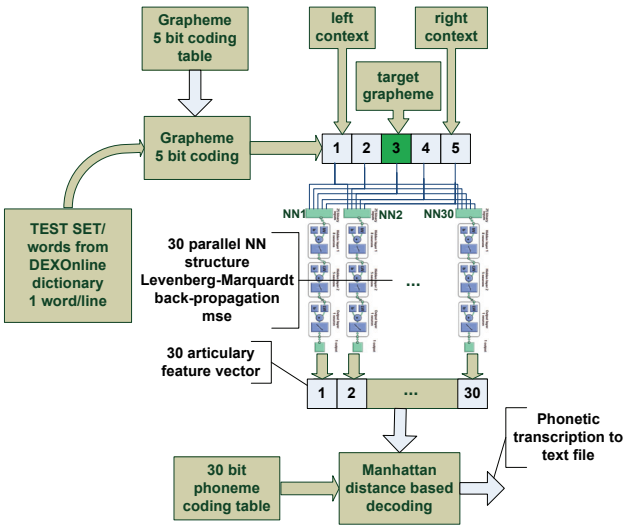

Fig. 2. ANN system architecture for automatic grapheme to phoneme conversion [15]

Fig 2. presents the architecture of our automatic grapheme to phoneme conversion system which is based on a parallel structure having 30 neural networks with 25 common inputs, each of them designed to detect the presence of an articulatory feature from the 30 features used to encode the Romanian language phonemes. Each network has to point out the presence or the absence of that feature at the network output. The words that are intended to be transcribed are read from a text file edited with one word by line and are presented at the input of each neural network. The beginning and the end of each word is appended with 2 white space characters (#) and after this, the input words are split into 5 character long sequences and binary coded using a 5 bit binary code. Always the central grapheme from the sequence of 5 graphemes is analyzed the other graphemes represent context information (2 graphemes for left context and 2 graphemes for right context). Therefore at the input of each neural network we have 5x5=25 bit information; hence we can deduce the number of network inputs. The words at the input of the system are shifted character by character until all component graphemes are presented to the input [15].

Even in this way the pronunciation dictionary development is a time consuming task, but the role of the human in this process is only to supervise and correct eventual errors of the system.

The created dictionary was exported in text format with UTF-8 character encoding and it is available in 2 different formats:

- HTK dictionary format with the following syntax for each line:
  <transcription><space><pronunciation>

- Festival dictionary format with the following layout for each line:
  ("<transcription>" <nil> (<pronunciation>))

Each phoneme is delimited by spaces in *<pronunciation>*.

Our system was designed to allow a speaker fluent in the target language to easily develop a pronunciation dictionary without having expert linguistic knowledge or advanced programming skills.

4. TESTING THE DICTIONARY

To achieve the best transcription results we have over-tested the pronunciation dictionary using Dictionary Maker.

We have recorded and segmented the Romanian phonemes for the used phoneme set, using Audacity, the free, cross-platform sound editor application [19]. The utterances were recorded by a young male speaker, fluent in Romanian language, the first author of this paper. These audio files were provided to Dictionary Maker in order to be used for generation of the sounded version for each transcription for the words included in the word list.

The system runs through the word list word by word, predicts a pronunciation and sounds out the phonemes of the word. The user listen to the generated pronunciation variant and provides a verdict with consideration to the accuracy of the word – pronunciation pair choosing one of the answers, according to Dictionary Maker tutorial [16]:

- Correct – The word is a valid word in the language concerned and its pronunciation as displayed is correct;
- Invalid – The word is not a valid word (e.g. it is a Universal Resource Locator, an email address), or it is spelled wrong, or it is only part of a word;
- Uncertain – The user is unable to decide whether the word and its pronunciation are valid;
- Ambiguous – There are multiple valid pronunciations;
- Proper noun – The word is a proper noun;
- Foreign – The word is a valid word from a foreign language, but not a word in the source language.

According to [5] with a number of 10k training words the grapheme level accuracy of the DEC algorithm is around 97%. In the condition the word level accuracy is given around 75%. If system is trained using 5k training words we have almost the same results.

5. CONCLUSIONS

We have created the first 100k+ words machine-readable Romanian language pronunciation dictionary based on the words from the *lexem* table of Dex.

The generated transcription dictionary together with the used grapheme and phoneme sets and the recorded and segmented audio samples for the used phoneme set can be freely downloaded from the NaviRO project website (http://users.utcluj.ro/~jdomokos/naviro/), and the use of this dictionary for any research purpose is completely unrestricted.

We appreciate that the results are very useful for Romanian language large vocabulary speech recognition system and text-to-speech system development.

The authors cannot guarantee the accuracy of the dictionary, nor its suitability for any specific purpose. In fact, we expect some errors, omissions and inconsistencies to remain in the dictionary however the entries were manually checked. Any suggestions, corrections and observations are welcomed. We intend to continually update the dictionary by correcting existing entries and by adding new ones. From time to time a new version will be posted on the project website.

As future work we can mention that our final goal is to generate a 1 million word pronunciation dictionary based on the inflected forms from the DexOnline dictionary so we welcome input from users in order to increase dictionary size and cover pronunciation variants. This 100k words dictionary will be a good input for Dictionary Maker in order to perform grapheme to phoneme mapping for the inflected words. Once all the inflected forms are transcribed, the resulting dictionary will be ready to be used for continuous speech phonetic transcription generation.

We are also interested to generate pronunciations for Romanian person names and institutions names because the needed word lists can be easily collected from the Internet. All these resources will be publicly available on the project website.

6. ACKNOWLEDGMENTS

This paper was supported by the project "Develop and support multidisciplinary postdoctoral programs in primordial technical areas of national strategy of the research - development - innovation" 4D-POSTDOC, contract nr. POSDRU/89/1.5/S/52603, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

Thanks for Marelie Davel and Etienne Barnard for Dictionary Maker application.

Thanks for Audacity developers.

7. 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] M. Bisani, H. Ney, "Joint-Sequence Models for Grapheme-to-

Phoneme Conversion", Speech Communication, Vol. 50, Elsevier, pp. 434–451, 2008.
[3] J.A. Gómez, M.J. Castro, "Automatic Segmentation of Speech at the Phonetic Level", Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, pp. 672-680, Springer-Verlag, London, UK, 2002.

[4] M. Davel, E. Barnard, "Pronunciation Prediction with Default&Refine", Computer Speech and Language, Vol. 22, Elsevier, pp. 374-393, 2008.

[5] M. Davel, E. Barnard, "Bootstrapping in Language Resource Generation", Proceedings of the 13th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pp. 97-100, Langebaan, South Africa, 2003.

[6] 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.
[7] Carnegie Mellon University (CMU) Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] British English Example Pronunciations (BEEP), http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

[9] 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.

[10] D. Jitcă, V. Apopei, F. Grigoras, "An Ann-Based Method to Improve the Phonetic Transcription Module of a TTS System for the Romanian Language", CD-ROM Proc. of the European Conference on Intelligent Technologies - ECIT 2002, Iasi, Romania, 2002.

[11] D. Jitca, H.-N. L. 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", Proc. of the 2nd Speech Technology and Human-Computer Dialogue Conference - SpeD, Bucharest, Romania, pp. 43-50, 2003.

[12] Ş.-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.

[13] M. A. Ordean, A. Şaupe, 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), Timişoara, Romania, pp. 401-406, 2009.

[14] A. Stan, J. Yamagishia, S. King and M. Aylettc, "The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate" Speech Communication, Vol 53, Issue 3, Elsevier, pp. 442-450, 2010.

[15] 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, Braşov Romania, 2011.

[16] Dictionary Maker application homepage on SourceForge: http://dictionarymaker.sourceforge.net/

[17] DexOnline - Transpunerea pe Internet a Unor Dicţionare de Prestigiu ale Limbii Române, http://dexonline.ro/

[18] Institutul de Lingvistică „Iorgu Iordan - Alexandru Rosetti" al Academiei Române, "DOOM - Dicţionarul Ortografic, Ortoepic şi Morfologic al Limbii Române (Editia a II-a, revizuita şi adăugită)", Editura Univers Enciclopedic, Bucureşti, 2005.

[19] Audacity application homepage on SourceForge: http://audacity.sourceforge.net/