MaRePhoR – An Open Access Machine-Readable Phonetic Dictionary for Romanian

Ştefan-Adrian Toma[1], Adriana Stan[2], Mihai-Lică Pura[1], Traian Bârsan[1]
[1]Computer Science Department, Military Technical Academy, Bucharest, Romania
[2]Communications Department, Technical University of Cluj-Napoca, Romania
stefan.toma@mta.ro, adriana.stan@com.utcluj.ro, puramihai@gmail.com

Abstract— **This paper introduces a novel open access resource, the machine-readable phonetic dictionary for Romanian - MaRePhoR. It contains over 70,000 word entries, and their manually performed phonetic transcription. The paper describes the dictionary format and statistics, as well as an initial use of the phonetic transcription entries by building a grapheme to phoneme converter based on decision trees. Various training strategies were tested enabling the correct selection of a final setup for our predictor. The best results showed that using the dictionary as training data, an accuracy of over 99% can be achieved.**

Keywords— phonetic dictionary, phoneme, grapheme, phonetic transcription, open access

I. INTRODUCTION

Wikipedia and the KTH Royal Institute of Technology recently announced that they plan to build the first crowd sourced speech engine [1] and transform its textual content into spoken data. If their initial efforts are targeted towards mainstream languages, their plan is to include all of the 280 languages in which Wikipedia is available. However, around 80% percent of these languages lack the freely available resources for such an engine to be developed. And among these languages, Romanian is also to be found. The lack of these resources has a major impact in the development of any natural language or speech processing algorithm or tool, and not just Wikipedia's endeavors. For example in Romanian, where the syllabification has 9 basic rules, a large set of the exceptions from these rules come from the need to distinguish between diphthongs and hiatus clusters--a grapheme-to-phoneme conversion problem which would require a phonetic dictionary. And of course, the need for any open resources is even more important in high-level applications, such as text-to-speech synthesis (TTS) or automatic speech recognition (ASR) systems. In [2], the authors present an extensive overview of the tools and resources available for Romanian, and conclude that "resources for the Romanian language are less represented than instruments, although they are essential for testing the designed tools." Since the article's publication there have been a number of additional open resources released, however their number is not significant in comparison to their necessity.

In aid of this lack of resources, we decided to create an open digital lexicon with a large number of entries, which will hopefully help, and also boost the research of the Romanian language technologies. The dictionary includes manual phonetic transcription of 72,375 words, performed by 4 experts. The selected list of words is a merger between the Romanian Scrabble Associations' official entries and the 15,517 words lexicon from [3].

This dictionary is not the first of its kind. However, it is the only open access phonetic dictionary with such a large number of manual entries. Other such resources include the one presented in [4], where Domokos et al. introduce a large phonetic dictionary (over 100,000 words). This dictionary is automatically generated from a smaller set of 5,000 words which were manually transcribed. In [5], Cucu et al. use a 600K words lexicon; however this dictionary is not available in an open access manner. Toma et al. [3] use a 15,517 words dictionary for training letter-to-sound (LTS) systems and a 4,779 words dictionary for testing.

For other languages, this type of resources has been better represented for some time. For example, for German there is the Bonn Machine-Readable Pronunciation Dictionary (BOMP) [6] and for English the Carnegie Mellon University (CMU) Pronouncing Dictionary [7], or the British English Example Pronunciations (BEEP) [8].

To determine the lexicon's effectiveness, we developed and tested grapheme-to-phoneme systems based on neural networks and decision trees. The results show that the accuracy is above the one obtained with previously developed smaller subsets of the lexicon, and to our knowledge, it is better than any other previously published study. Furthermore, the results show that increasing the dictionary size above a threshold, does not improve the effectiveness of the automatic phonetic transcription giving rise to the question of how important is phonetic dictionary size.

The paper is structured as follows. Section 2 is dedicated to describing the phonetic dictionary, while in Section 3, phonetic transcription experiments are presented. In Section 4 we investigate the dictionary's size relevance. Conclusions and future work are presented in Section 5.

II. DICTIONARY DESCRIPTION

Creating a phonetic dictionary for Romanian is not a trivial task. Although it is considered mostly a phonetic language [9], the 31 letters in the Romanian alphabet are pronounced using 34 phonemes. Some letters have more than one pronunciation, some are pronounced using a combination of sounds and certain groups of letters are pronounced as only one sound. For

TABLE 1. THE PHONEMES OF THE ROMANIAN LANGUAGE

SAMPA	Graphic	Phonetic	SAMPA	Graphic	Phonetic
Vowels			**Consonants**		
i	P**I**N	p **i** n	g_j	**G**HEM	**g_j** e m
e	F**E**L	f **e** l	g	**G**ÂT	**g** 1 t
a	C**A**P	c **a** p	ts	**Ț**ARĂ	**ts** a r @
@	M**Ă**R	m **@** r	v	**V**ALE	**v** a l e
o	L**O**C	l **o** k	s	**S**AC	**s** a k
u	S**U**R	s **u** r	z	**Z**I	**Z** i
1	F**Â**N	f **1** n	S	**Ș**A	**S** a
Semivowels			Z	**J**OI	**Z** o j
j	DO**I**	d o **j**	h	**H**AM	**h** a m
e_X	N**E**A	n **e_X** a	m	**M**ARE	**m** a r e
w	SA**U**	s a **w**	n	**N**AS	**n** a s
o_X	**O**ARE	**o_X** a r e	l	**L**APTE	**l** a p t e
Consonants			r	**R**ÂS	**r** 1 s
p	**P**ĂR	**p** @ r	k_j	**C**HEL	**k_j** e l
b	**B**AR	**b** a r	dZ	**GE**AM	**dZ** a m
t	**T**UN	**t** u n	f	**F**ATĂ	**f** a t @
d	**D**AR	**d** a r	**Non-syllabic vowel**		
k	**C**AI	**c** a i	i_0	LUP**I**	l u p **i_0**
tS	**CE**AS	**tS** a s			

diphthongs, tripthongs and hiatus there is no clear rule, other than syllabication and lexical stress. However, more than half the letters of the Romanian alphabet are pronounced using only one class of sounds for each letter. The complete list of phonetic symbols for the Romanian language along with pronunciation examples is presented in Table 1 and adheres to the Speech Assessment Method Phonetic Alphabet (SAMPA)[10].

Although most of the Romanian graphemes are pronounced with a single phonetic sound, the exceptions still require a large dataset of manually transcribed inputs for a highly accurate automatic phonetic transcription, and the Machine-Readable Phonetic Dictionary for Romanian MaRePhoR dictionary introduced in this paper, aims to enable this objective.

The dictionary consists of 72,375 words with a total of 591,570 letters. The entries are words from the Romanian Scrabble Association official list of words and from a 15,517 words dictionary developed according to the SpeechDat specifications [11]. The transcriptions of the word entries were obtained in a two step process. First, the existing 15,517 words dictionary from [3] was used to train a neural network-based grapheme-to-phoneme converter, which in turn provided the semi-supervised transcriptions for our 72,375 entries. These transcriptions were afterwards manually verified and corrected, if necessary, based on the pronunciation rules found in [12] and [13].

One of the reasons why we chose to use a different list of words from the one present in NaviRO [4] is the fact that for this list we also have the lexical stress and syllabication of the entries, as performed in [14]. However, this information is not yet present in our release due to the fact that we were unable to manually check it entirely. Nonetheless, this allows for further development of the dictionary since syllabication, lexical stress and pronunciation in Romanian are interconnected. There are some cases in which the phonetic transcription is completely based on the syllabication or lexical stress positioning. For example, in Romanian the vowels *E*, *I*, *O* and *U* generate most of the errors in phonetic transcriptions. Either in conjunction with other vowels or with themselves, they can create diphthongs, tripthongs or hiatus, depending on lexical stress and syllabication. In MaRePhoR there are 16,235 words that contain hiatus and 7,136 words with possible diphthongs and tripthongs.

TABLE 2. OCCURRENCES OF LETTERS AND LETTER SEQUENCES PRONOUNCED AS ONE SOUND

Letter sequence	No. of words
GHE	254
GHI	323
CHE	598
CHI	998
CE	2,585
CI	3,414
GE	1,438
GI	1,707

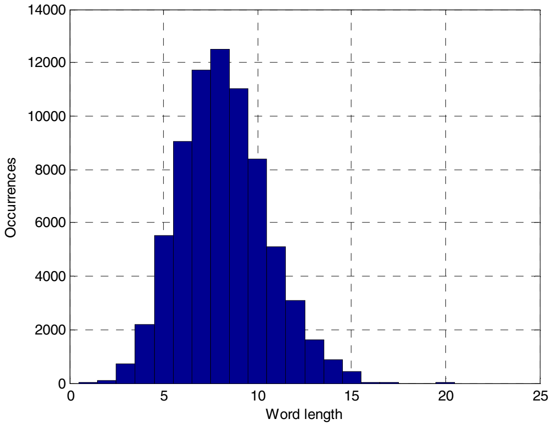

Figure 1: *Word length (number of letters) histogram of the MaRePhor entries*

Other problematic letters are *X* with 1,470 occurrences and B with 9,758 occurrences. The letter *X* is pronounced as a sequence of two sounds (i.e. *ks* or *gz* as in *PIX p i ks* and *EXEMPLU e gz e m p l u*), while *B* is uttered either as *p* or as *b* (e.g. *BIROU b i r o w* and *SUBȚIRE s u p ts i r e*). There are also certain letter sequences in Romanian which are pronounced as one sound. These are *HE,* and *HI* when preceded by the letters *G* or *C* and *CE, CI, GE* and *GI.* The sequences' occurrence in MaRePhoR is given in Table 2.

Of the 72,375 words in MaRePhoR, 128 are neologisms, and have different pronunciation rules than the standard Romanian ones. The length of the word entries varied from 1 to 25 letters and is distributed as shown in Figure 1.

The dictionary entries are represented in ASCII. To enable the use of diacritics we made the following convention: letters which are present in the ASCII set are written in uppercase, while the diacritics use lowercase symbols. The phonetic transcription uses blank spaces to separate the phonetic symbols.

TABLE 3. LETTER AND PHONE OCCURRENCE STATISTICS WITHIN MAREPHOR

Letter	No. of occurrences		Phone(s)	No. of occurrences	
	No.	%		No.	%
A	49,078	9.5963	a	49,080	9.5918
Ă	17,792	3.4785	@	17,793	3.4773
Â	2,372	0.4638	1	2,361	0.4614
B	8,630	1.6873	b	8,490	1.6592
			p	137	0.0268
C	27,534	5.3832	k	20,813	4.0675
			k_j	1,351	0.2640
			tS	5,308	1.0374
D	11,962	2.3387	d	11,962	2.3378
E	50,224	9.8193	e	48,014	9.3835
			e_X	1,807	0.3531
			ee_X	11	0.0021
			je	24	0.0047
			=	363	0.0709
F	7,537	1.4736	f	7,538	1.4732
G	9,231	1.8048	g	5,964	1.1656
			g_j	435	0.0850
			dZ	2,830	0.5531
H	4,183	0.8178	h	2,375	0.4642
			=	1,801	0.3520
I	53,382	10.4368	i	42,974	8.3985
			i_0	838	0.1638
			ij	6,762	1.3215
			j	3,246	0.6344
Î	1,958	0.3828	1	1,958	0.3827
J	1,774	0.3468	Z	1,773	0.3465
K	113	0.0221	k	48	0.0094
L	23,801	4.6533	l	23,791	4.6495
M	16,951	3.3141	m	16,946	3.3118
N	29,291	5.7267	n	29,291	5.7244
O	33,317	6.5138	o	32,218	6.2964
			o_X	1,098	0.2146
P	15,191	2.9700	p	15,174	2.9655
Q	10	0.0020	k	9	0.0018
R	45,367	8.8697	r	45,367	8.8662
S	18,637	3.6437	s	18,616	3.6382
Ş	4,241	0.8292	S	4,245	0.8296
T	38,636	7.5538	t	38,622	7.5480
Ţ	5,490	1.0734	ts	5,499	1.0747
U	20,504	4.0088	u	19,366	3.7847
			w	789	0.1542
			uw	351	0.0686
V	5,955	1.1643	v	5,946	1.1620
W	14	0.0027	v	9	0.0018
X	1,331	0.2602	ks	1,203	0.2351
			gz	128	0.0250
Y	25	0.0049	i	18	0.0035
Z	6,950	1.3588	z	6,945	1.3573

In order to have a synchronization between word entries and their transcriptions (i.e. have the same number of symbols in both), we used the "=" symbol to represent letters which are not pronounced (e.g. *CEAS ts = a s*). If a letter is pronounced by two or more phones, they are not separated by blank spaces (e.g. *EXEMPLU e gz e m p l u*). The dictionary is also available without making use of the above convention.

In Table 3 we show a complete overview of the number of occurrences for each letter and phone or phone group within the MaRePhoR dictionary.

MaRePhoR is available at http://speech.utcluj.ro/marephor/ and http://www.mta.ro/marephor/ under a Creative Commons Attribution Non-Commercial 3.0 Unported (CC BY-NC) license [15].

III. PHONETIC TRANSCRIPTION EXPERIMENTS

To test the dictionary's viability as a dependable resource for higher-level applications, such as automatic speech recognition or text-to-speech systems, we used it as training data in a grapheme-to-phoneme converter.

Various algorithms and training strategies were employed. However, the feature building strategy remains the same across all experiments. That is, the current letter is combined with a left and right context window, and the window does not span over preceding or succeeding words, as there is no evidence that in Romanian the pronunciation of a word is influenced by the neighboring ones. If the window is larger than the word's length, the context is padded with a non-letter symbol, in this case, the symbol '-'.

Regarding the algorithms, we opted for those which would not only have the smaller error rate, but would also provide an efficient, real-time prediction. Therefore, we selected four such algorithms, already available in Weka [16 and 17]: decision trees, decision table, attribute selected classifier and multilayer perceptron. The accuracy of these algorithms when evaluated in a 10-fold cross validation manner is shown in Table 4. It can be noticed that the decision trees have the smallest error rate. As a result, we used them in the next experiments.

The accuracy of the decision tree based G2P converter was tested for different sizes of the training and testing datasets, as well as window lengths. Results are presented in Table 5.

The Romanian language's letter-to-sound rules are quite straightforward with most of the letters having a one-to-one correspondence with a phoneme. Out of the 31 letters in the Romanian alphabet, only a few have a context dependent pronunciation and these are: *B*, *C*, *E*, *G*, *H*, *I*, *O*, *U* and *X*. We will refer to this subset, as the *problematic letters*. We did not include in this pronunciation variation list any letter which has a different pronunciation due to the neologism nature of the origin word, such as *K*, *Q*, *Y* or *W*.

It is possible that for some of the least frequent letters, the training data might be insufficient. Hence, we tested this hypothesis and built both individual decision trees for each problematic letter, as well as a global tree for all of them. The results are shown in Table 6 and Table 7. The window length was set to 5 letters, and the results are for the 10-fold cross-validation accuracy metric. For most of the letters, the best results are obtained by the global tree, with the notable exception for the letter *X*, where the individual tree has a 7% relative increase in the accuracy.

TABLE 4. ACCURACY [%] OF VARIOUS ALGORITHMS FOR THE GRAPHEME-TO-PHONEME CONVERTION FOR A WINDOW LENGTH OF 5 LETTERS AND PREDICTING ONLY THE PROBLEMATIC LETTERS

Algorithm	Accuracy [%]
Decision Table	98.48
Decision Tree	**99.18**
Attribute Selected Classifier	98.69
Multilayer Perceptron	98.87

TABLE 5. ACCURACY [%] OF THE DECISION TREES TRAINED TO PREDICT THE ENTIRE LETTER SET (591,570 FEATURES)

Train-test split	Window length		
	3	5	7
1 – 99	**98.69**	98.63	98.12
25 – 75	99.43	**99.49**	99.42
50 – 50	99.44	**99.57**	99.50
75 – 25	99.46	**99.59**	99.53
100	99.45	**99.61**	99.55

TABLE 6. ACCURACY [%] OF THE DECISION TREES TRAINED TO PREDICT ONLY THE PROBLEMATIC LETTERS (243,882 FEATURES)

Train-test split	Window length		
	3	5	7
1 – 99	**97.14**	96.79	96.86
25 – 75	98.66	**98.82**	98.62
50 – 50	98.70	**98.97**	98.84
75 – 25	98.69	**99.02**	98.96
100%	98.72	**99.11**	98.99

TABLE 7. ACCURACY OF THE INDIVIDUAL DECISION TREES TRAINED FOR EACH **PROBLEMATIC LETTER,** VERSUS THE ACCURACY OF THE PREDICTION FOR THE RESPECTIVE LETTER WHEN USING THE DECISION TREE TRAINED ON **ALL THE LETTERS.**

Letter	Tree type	
	Individual tree	Global tree
B	99.82	**99.85**
C	**99.92**	99.91
E	99.68	**99.69**
G	99.78	**99.93**
H	99.37	**99.39**
I	**97.60**	97.58
O	99.84	**99.87**
U	98.91	**98.94**
X	**98.09**	97.89

Because randomly splitting the dictionary into training and testing sets would not necessarily yield an accurate evaluation of it being used as training data, we wanted to put it to test in another scenario as well. For this we wanted to see the performance of both the dictionary and the decision trees for predicting the most common words in Romanian. The word list contains 4,779 entries selected from 93 Romanian literary and scientific works, and was collected in [18]. Out of the 4,779 entries, MaRePhoR contains 65%.

The letter error rate is **0.279%** while the word error rate is **1.82%**. Confusion matrices are presented in Table 8 through Table 12, for the problematic letters.

It is worth noting that the letters *O*, *C*, *G* and *H* are perfectly transcribed. The letter I has the highest number of errors (61 from a total of 88). More than half of these errors are generated by the final *I* in infinitive verbs and some nouns and adjectives. Part of speech and morphological information is important in deciding the correct pronunciation of the final *I*, and it will be made available in future developments of the dictionary.

Most detected errors do not have an impact on speech intelligibility. However, some words are homographs and heteronyms. For most of them the pronunciation generated by the converter was the one in the dictionary. However, for three words valid pronunciations, but different from the ones in the dictionary were generated. These were considered errors.

TABLE 8. CONFUSION MATRIX FOR THE LETTER E

E	**e**	**e_X**	**e e_X**	**je**	**=**
e		2			
e_X	2				
ee_X	1				
je	7				
=					

TABLE 9. CONFUSION MATRIX FOR THE LETTER I

I	**i**	**j**	**ij**	**=**	**i_0**
i		3		1	15
j	13		6		
ij	2	3			
=			1		
i_0	17				

TABLE 10. CONFUSION MATRIX FOR THE LETTER U

U	**u**	**w**	**uw**	**=**
u		1		
w	2			
uw		5		
=				

TABLE 11. CONFUSION MATRIX FOR THE LETTER X

X	**ks**	**gz**
ks		
gz	3	

TABLE 12. CONFUSION MATRIX FOR THE LETTER B

B	p	b
b		
p		4

IV. DOES SIZE MATTER ?

The general consensus is that a bigger phonetic dictionary is better. But given the resources needed to create phonetic dictionaries verified by experts, one has to ask how many words should it contain? The answer generally depends on usage of the dictionary.

For training automatic speech recognition (ASR) engines, the phonetic dictionary has to contain at least the words used in the training set, so it depends on domain the ASR is used for. However, no one can guarantee that only the words used for training will be used in ASR or TTS applications.

Therefore the question becomes: which is the smallest dictionary from which one can generate accurate enough phonetic transcriptions? As shown in the previous section, increasing the dictionary size leads to an increase in the accuracy. We investigated this aspect in more detail. Using the 4,779 words dictionary we computed the LER obtained for classifiers trained using an increasing number of instances. We started from 5,000 instances and increased it. Prior to instance selection, we randomized the instance set. The results are depicted in Figure 2.

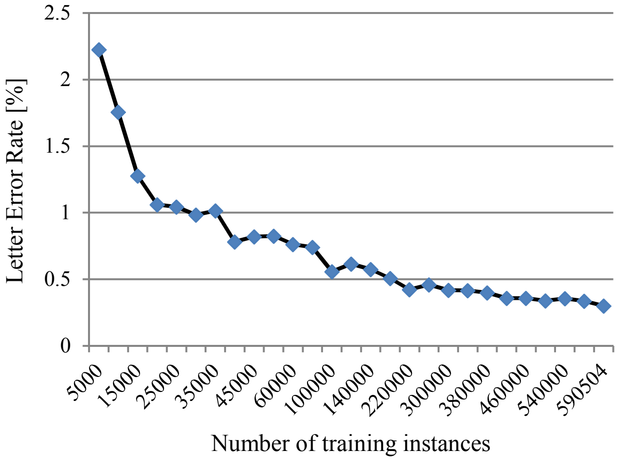

Figure 2: *Letter error rate vs. number of instances*

In Figure 2, the letter error rate decreases abruptly till the instance set reaches 25,000 entries. This roughly corresponds to 4,000 words (calculated for the mean value from Figure 1). Then it decreases at a much slower rate until at around 300,000 instances (37,500 words). After this point the letter error rate decreases very slowly. This suggests that increasing dictionary size doesn’t lead to significant increase in performance.

However, it is possible that adding more training instances for the problematic letters may lead to further improvement. Furthermore, some other grapheme to phoneme conversion method may lead to better results, further decreasing the LER. Hence, further investigation is required.

V. CONCLUSIONS AND FUTURE WORK

The paper introduced a new open access phonetic dictionary for Romanian, called MaRePhoR. The dictionary is the largest resource of this kind, and was developed with the hope that it will draw more attention towards developing new, easily accessible, high quality computer interaction systems for Romanian.

We tested the use of the dictionary for building phonetic transcribers for Romanian, and evaluated several algorithms and training strategies. The best results showed an accuracy of 99.61%, similar to the results in [19] and [21]. However, a direct comparison to other letter to sound systems is difficult to make since the experimental conditions were not the same and it is beyond the scope of this paper.

As future work, we would like to extend the dictionary, and include the syllabication and accent positioning for all its entries. We are also planning on adding the information to the online Romanian dictionary Dexonline [20]. Within this platform, the resource’s visibility and high number of viewings will also help us correct any potential errors we might have missed, as well as adding new entries provided by the users.

ACKNOWLEDGMENT

The research leading to these results was partially funded by the Romanian Ministry of Education, under the grant agreement PN-II-PT-PCCA-2013-4 No. 6/2014 (SWARA).

REFERENCES

[1] KTH and Wikipedia develop first crowdsourced speech engine. 10th May 2016. Internet: https://www.kth.se/en/forskning/artiklar/kth-hjalper-wikipedia-borja-prata-1.631897. [30.03.2016]

[2] Trandabăț, D., Irimia, E., Mititelu, V. Barbu, Cristea, D., and Tufiş, D. (2012). Limba română în era digitală – The Romanian Language in the Digital Age. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

[3] Ş A. Toma, T. Bîrsan, F. Totir and E. Oancea, "On letter to sound conversion for Romanian: A comparison of five algorithms," Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on, Cluj-Napoca, 2013, pp. 1-6.

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

[5] H. Cucu, L. Besacier, C. Burileanu, A. Buzo, “Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods,” 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.260-265, 11-15 Dec. 2011.

[6] T. Portele, J. Krämer, D. Stock. "Symbolverarbeitung im Sprachsynthesesystem Hadifix" Proc. Elektronische Sprachsignalverabeitung 1995, Wolfenbüttel, 97-104.

[7] Carnegie Mellon University (CMU) Pronouncing Dictionary, Internet: http://www.speech.cs.cmu.edu/cgi-bin/cmudict [03.30.2016]

[8] British English Example Pronunciations (BEEP), Internet: http://svr-www.eng.cam.ac.uk/comp.speech/ Section1/Lexical/beep.html [03.30.2016]

[9] I. Calotă, Mică enciclopedie a românei corecte. Niculescu Publishing house, Bucharest, Romania, 2001.

[10] Speech Assessment Methods Phonetic Alphabet – SAMPA. Internet: http://www.phon.ucl.ac.uk/home/sampa/romanian.htm [30.03.2016].

[11] [SpeechDat] "The SpeechDat project". Internet: http://www.speechdat.org [05.01.2011].

[12] Academia Româna (Romanian Academy), Dicţionarul ortografic, ortoepic şi morfologic al limbii române, Ediţia a doua (Romanian language orthographic, orthoepic and morfologic dictionary, the second edition), Univers Enciclopedic, Bucharest: Romania, 2005.

[13] F. Şuteu and E. Şoşa, Dicţionar ortografic al limbii române (Eng. Romanian language spelling dictionary), Vestala Publishing house, Bucharest, Romania, 1993.

[14] S. A. Toma, E. Oancea and D. P. Munteanu, "Automatic rule-based syllabication for Romanian," Speech Technology and Human-Computer Dialogue, 2009. SpeD '09. Proceedings of the 5-th Conference on, Constant, 2009, pp. 1-6.

[15] "Creative commons — Attribution-NonCommercial 3.0 Unported — CC BY-NC 3.0," in Creative commons — Attribution-NonCommercial 3.0 Unported — CC BY-NC 3.0. [Online]. Available: https://creativecommons.org/licenses/by-nc/3.0/. Accessed: Feb. 23, 2017.

[16] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[17] M. Hall et al., "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Vol. 11, Issue 1, pp. 10-18, 2009.

[18] A. Vlad, A. Mitrea, and M. Mitrea. Limba română scrisă, ca sursă de informaţie. (Eng. Written Romanian Language as a Source of Information). Paideia Publishing House. 2003.

[19] M. Stănescu (Paşca), H. Cucu, A. Buzo and C. Burileanu, "ASR for low-resourced languages: Building a phonetically balanced Romanian speech corpus," 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, 2012, pp. 2060-2064.

[20] "Dexonline," in Dexonline - Dicţionare ale limbii române. [Online]. Available: https://dexonline.ro/. Accessed: Feb. 23, 2017.

[21] Tiberiu Boroş. "A Unified Lexical Processing Framework Based on the Margin Infused Relaxed Algorithm. A Case Study on the Romanian Language," in Proceedings of the International Conference Recent Advances in Natural Language Processing. Hissar, Bulgaria, 2013.