

## A ROBUST DIACRITICS RESTORATION SYSTEM USING UNRELIABLE RAW TEXT DATA

Lucian Petrică\*, Horia Cucu, Andi Buzo, and Corneliu Burileanu

University Politehnica of Bucharest, Speech and Dialogue Research Laboratory, Bucharest, Romania

\*Corresponding author (E-mail: [lucian.petrica@upb.ro](mailto:lucian.petrica@upb.ro))

### ABSTRACT

Statistical language models are utilized in many speech processing algorithms, e.g., automatic speech recognition (ASR). Such a model is created from a text corpus, but many of the text corpora for Romanian are unreliable with respect to the use of diacritic marks, i.e., diacritics are either partially or completely missing, resulting in low quality language models. We present a methodology for restoring diacritic marks to an unreliable text corpus, which requires no text resources apart from the corpus itself. The proposed methodology (i) identifies sections of the input corpus which are correct with respect to the use of diacritics, (ii) utilizes these sections to train a diacritics restoration system (DRS), and (iii) utilizes the DRS to correct the remaining sections of the corpus. We compare the DRS trained at (ii) with state-of-the-art systems, and observe up to 12% improvement with regard to the correctness of diacritic restoration. Furthermore, we utilize our methodology to create improved language models for the ASR system developed by the Speed laboratory, and demonstrate a decrease of 14% in perplexity and a 20% reduction of the out-of-vocabulary rate as a result.

**Index Terms**— Diacritics, speech recognition

### 1. INTRODUCTION

The Romanian language utilizes three diacritic marks, and has five characters with diacritics: *ă*, *â*, *î*, *ș*, and *ț*. The diacritic marks modify the pronunciation of the four base characters: *a*, *i*, *s*, and *t*. Despite the small number of diacritic marks, as much as 40% of the words in a Romanian text utilize diacritics [1]. Missing diacritics in a text can lead to grammatically incorrect words, or ambiguous words for which a missing diacritic mark modifies meaning but maintains grammatical correctness. A diacritics restoration system (DRS) parses an input text and replaces words with missing diacritics with their correct versions. When multiple grammatically correct replacements are possible, the ambiguity is resolved using an *n-gram* [2] language model, which assigns a probability to each replacement candidate based on the previous *n* corrected words. Creating this language model is called *training* the DRS.

Unfortunately, the Romanian language is a so-called under-resourced language, for which there is insufficient availability of quality text corpora from reliable sources, and

for which most resources have to be acquired online [3, 4]. Almost all text resources which can be acquired online for the Romanian language have unreliable, if any, use of diacritics. The unreliability of text resources is problematic for statistical systems which utilize *n-gram* language models, including DRS and automatic speech recognition (ASR) systems. If a diacritics restoration system (DRS) is available, it may be utilized to correct the available text resources, therefore increasing ASR quality. However, a DRS must itself be trained utilizing reliable text resources, which may not be available. Therefore, the research question is how to break this circular dependency by training a high-performance DRS utilizing an unreliable text corpus.

This paper proposes a methodology whereby an initially unreliable text corpus is filtered in order to identify sections of the corpus with high likelihood of correctness with regard to diacritics use. These sections may be used by themselves or in conjunction with reliable (e.g., hand-corrected) resources to train a DRS of better quality than can be obtained using reliable resources alone. The resulting DRS is used to correct the previously filtered-out sections of the text corpus, which are subsequently considered reliable. We use this enhanced corpus to train an ASR system, and demonstrate increases in the quality of the ASR output text. The proposed methodology requires no reliable text resources, but is able to utilize such resources when available.

The remainder of this paper is structured as follows. Section 2 presents previous approaches to diacritics restoration in the context of speech processing. Section 3 presents the methodology for diacritics restoration in detail. In Section 4, we present the experimental set-up and results of our evaluation of the proposed methodology, while Section 5 lists conclusions and avenues for future work.

### 2. RELATED WORK

Several methods have been proposed for diacritics restoration of Romanian language text. A knowledge-based method of restoration is utilized in [5] to make decisions in ambiguous situations, resulting in 2.25% diacritics word error rate (WER) and 0.60% diacritics character error rate (ChER). Systems based on sequential filtering using word-suffix *n*-grams, for use in speech synthesis, were reported in [6] and [7], and demonstrate a best result of 1.4% WER and

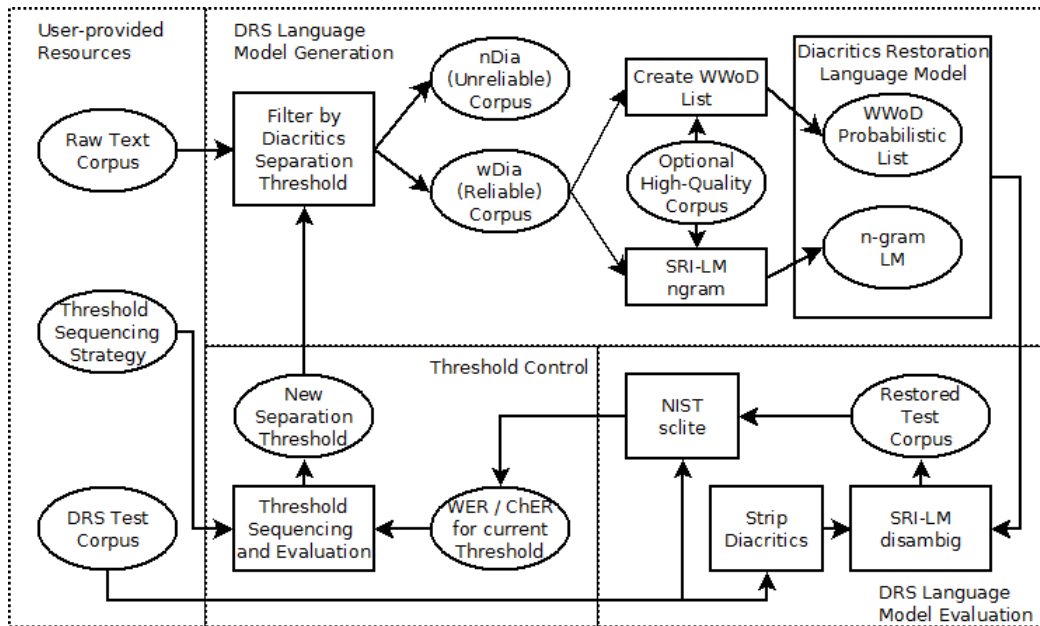


Figure 1: DRS Training Methodology

0.4% ChER. Experiments in [8] utilizing n-gram language models and probabilistic maps, trained with a corpus with known-good diacritics, resulted in a 1.99% WER and 0.48% ChER for a 3-gram probabilistic language model.

Further work in [9] extends the findings presented in [8] with evaluations of diacritics restoration in the context of ASR. The authors utilize the restoration system developed in [8] to increase ASR quality in one of two ways. The first method is to train the ASR with the acquired unreliable text corpus, in which case the output text may also have missing diacritics, and diacritic restoration is performed on the output. Alternatively, diacritic restoration is applied to the ASR train corpus, thereby increasing the quality of the train corpus and the likelihood that the system will output text with correct diacritics. The authors report a total WER at the ASR output of 30.5% and 29.7% respectively for the two methods, an improvement from the 64.5% WER of the non-DRS-augmented ASR system.

It is obvious that diacritics restoration can improve the output of ASR, and it has been experimentally proven that correcting the LM training corpus provides better results, using the same DRS, than diacritics restoration on ASR output. We also note that all statistical diacritics restoration systems presented in previous work utilize reliable, often hand-corrected text corpora for training. We propose that a DRS can be trained utilizing unreliable text corpora, while maintaining restoration performance. Our method and experiments will therefore focus on matching or improving the performance of previous work on diacritics restoration and automatic speech recognition, utilizing unreliable text resources acquired online for training both the diacritics restoration language model and the automatic speech recognition language model.

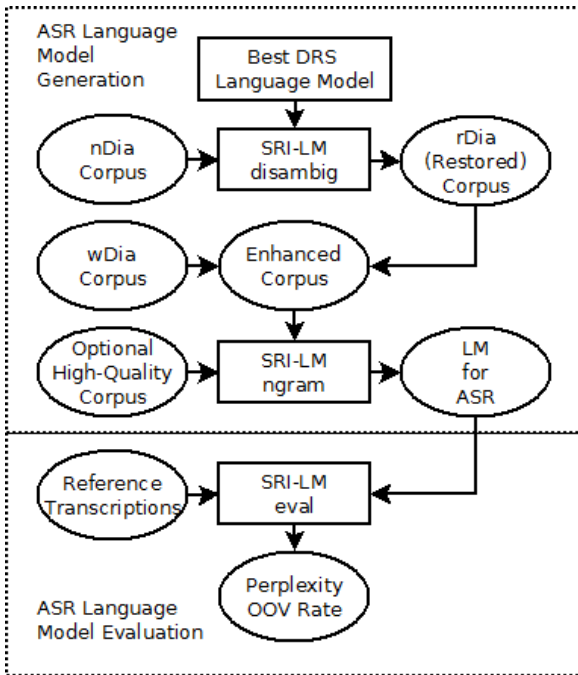
### 3. DIACRITICS RESTORATION METHODOLOGY

The proposed methodology is designed to generate diacritic restoration language models from text corpora acquired online (e.g., news articles), for use in automatic speech recognition. Functionally, the process can be divided into a DRS training step, illustrated in Figure 1, and ASR enhancement, illustrated in Figure 2. Although the DRS is utilized for ASR enhancement in this paper, it can also be utilized for general-purpose diacritic restoration in Romanian texts.

#### 3.1. DRS Training

Figure 1 presents the processing steps required for DRS training. We assume a text corpus has been acquired online using a web crawler, and consists of a collection of files, whereby each file represents a news article, transcription or other form of text. The unreliable raw text corpus is first cleaned (not pictured) to ensure the standard diacritic marks are used throughout the text. This step is necessary because many Romanian texts utilize non-standard diacritic marks, which were in use in text editing software over time. Following this step, each file in the corpus is processed to determine its corresponding ratio  $R_d$ , expressed in Equation 1, where  $N_d$  is the number of characters with diacritics ( $\tilde{a}$ ,  $\hat{a}$ ,  $\hat{i}$ ,  $\tilde{s}$ , and  $\tilde{t}$ ) and  $N_b$  is the number of base characters (a, i, s, and t).

$$R_d = N_d / (N_d + N_b) \quad (1)$$

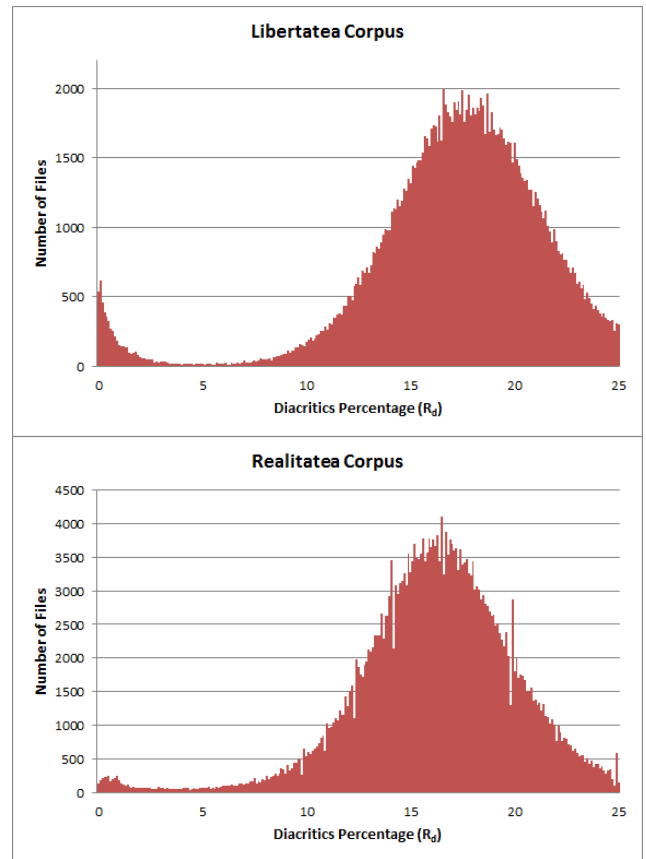


**Figure 2: ASR Training Methodology**

We assert that all files with a ratio below a given threshold  $T_d$  either are lacking diacritic marks, or contain significant errors, and therefore separate the raw text corpus into two corpora. The *wDia* corpus contains all files with ratios above the threshold, and is therefore considered the “high-quality” corpus, while the *nDia* corpus is considered low-quality and is not used in the DRS training process. The *wDia* corpus is used, by itself or in conjunction with other high-quality text corpora (not pictured), for the training of the DRS language model, using the ngram tool from the SRI Language Modeling Toolkit (<http://www-speech.sri.com/projects/srilm>).

The choice of filtering threshold is essential to the performance of the trained DRS, and therefore the methodology requires several thresholds to be utilized, resulting in several systems for diacritics restoration, of which only the best performing is kept. A threshold selection strategy must be defined, which may be as simple as cycling through a sequence of thresholds. In our implementation of this methodology, threshold selection is manual, but it may be performed automatically.

To assess the performance of a DRS, a hand-written, high-quality test corpus is used as reference. The diacritics are stripped from the test corpus, and restored by the DRS under evaluation, using SRI-LM disambig. The initial test corpus and the restored test corpus are compared using the *scite* tool from the NIST Speech Recognition Scoring Toolkit (<http://www.nist.gov/speech/tools>), which reports the word error rate and the character error rate. We assert that, given a sufficiently large test corpus, diacritic



**Figure 3: Online Corpora Diacritics Characteristics**

restoration performance on the test corpus is representative of the real-world performance of the DRS.

To utilize our proposed methodology, a user must provide an unreliable corpus of raw text data, a threshold selection strategy which determines which separation thresholds will be evaluated, and a small reliable text corpus for evaluation of the diacritics restoration system. This evaluation corpus may be relatively small compared to the unreliable text corpus.

### 3.2. ASR Enhancement

Figure 2 illustrates the process by which the best DRS identified is utilized to enhance the ASR system. Typically, only the *wDia* sub-corpus would be used for ASR training. Using the DRS, the diacritics are restored to the *nDia* corpus, resulting in the *rDia* corpus which is added to *wDia* and an optional reliable corpus to result in the final training corpus for ASR use. Because ASR quality is approximately proportional to the training corpus size, we expect the quality of ASR output to also increase with the addition of *rDia* to the training corpus. To verify this assertion, the ASR system is tested on standard recognition tasks, and quality metrics are recorded.

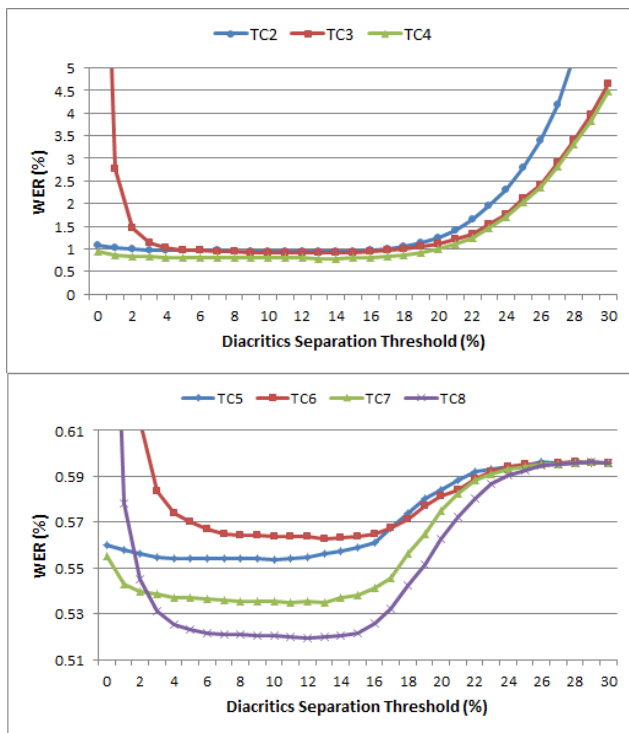


Figure 4: DRS Performance

#### 4. EVALUATION

The proposed methodology will be evaluated for diacritics restoration performance, but also in conjunction with automatic speech recognition, utilizing unreliable text resources.

##### 4.1. DRS Experimental Set-up and Results

For the purposes of testing the proposed methodology, three corpora have been acquired online, using web news outlets as source. Table 1 summarizes the characteristics of the corpora, with regard to number of text files and total number of words. For reference, we included the DRS used in [8], and also enhanced it with a known-good corpus, *Meetings*, which is hand-written. These two reference corpora are listed as *ref* and *ref-enhanced* respectively. The rightmost column lists the number of files in the corpus which lack diacritic characters completely, and are with high probability incorrect with regard to diacritics use.

Figure 3 presents two histograms of the *Realitatea* and *Libertatea* corpora, to illustrate how the corpus files are distributed with regards to  $R_d$ . As expected,  $R_d$  follows a normal distribution, but with both corpora containing an abnormally large number of files towards  $R_d=0$ . This is caused by files in both corpora which are only partially written using diacritics. As can be seen, the effect is more pronounced for the *Libertatea* corpus than for the *Realitatea*

corpus. The *Antena3* corpus (not pictured) exhibits this characteristic much less than the other online corpora. For reference, the *Meetings* corpus scores no lower than 15% on the same  $R_d$  metric.

For DRS evaluation, we utilize the test corpus *MeetingsTest*, consisting of 4 million words. We utilize each of the two reference corpora, *TC0* and *TC1* as described in Table 2, to train a DRS. We also apply our proposed DRS generation methodology (pictured in Figure 1) to each of the three online corpora, varying  $T_d$  in 1% increments, from 0 to 30%. For each generated DRS, we measure WER and ChER. The proposed methodology is applied to a total of seven corpora, TC2 to TC8 as described in Table 2, by combining the available text resources in multiple ways.

Table 1: Corpora Characteristics

Corpus Name	# of files	# of words	# of files w/o any diacritics
Ref	1	15M	0
Ref-enhanced	6K	55M	0
Antena3	137K	27M	6K
Realitatea	367K	68M	54K
Libertatea	325K	72M	153K

Table 2: DRS Training Corpora

Evaluation Corpus	Components
TC0	Ref
TC1	Ref-enhanced
TC2	Antena3
TC3	Libertatea
TC4	Realitatea
TC5	TC1, TC2
TC6	TC1, TC3
TC7	TC1, TC4
TC8	TC1, TC2, TC3, TC4

Figure 4 illustrates the observed variation in WER for the various train corpora when  $T_d$  is increased. *TC3* is the largest of the news corpora, but also contains the largest number of files with no diacritics. Consequently, its WER for threshold  $T_d = 0$  is high, at 13%. *TC2* and *TC4* also exhibit a slightly higher WER at threshold  $T_d = 0$ . For all three corpora, minimum WER is achieved in the 8% to 12% range. Larger diacritics separation thresholds induce an increase in WER due to the decrease in the size of the wDia corpus. *TC2*, which is the smallest of the three corpora, exhibits the most pronounced increase in WER. The WER performance of *TC5-TC8* varies less with the separation threshold, because these corpora are constructed by combining unreliable text resources with *TC1* (a high-quality corpus), ensuring at least a moderate performance will be achieved irrespective of the size and quality of wDia.

Table 3 presents a comparison of WER and ChER for *TC1-TC8*, listing for *TC5-TC8* only the best DRS obtained through our proposed methodology, and the diacritics separation threshold found to perform best. *TC1* provides a marked improvement over *TC0*, the reference DRS from [8]. *TC2* performs best when the diacritics separation threshold is at 11%, but does not improve over *TC1*. The restoration systems developed using our proposed methodology and only unreliable online resources perform 17% to 30% better than previous work, but worse when compared to the DRS trained with *TC1*, which is a large hand-written training corpus.

For *TC5* to *TC8*, which use both online and hand-written resources, restoration performance is very good. Most notably, the DRS generated using our methodology with *TC8* as training corpus is able to achieve a 54% improvement over [8], and 12% improvement over the DRS trained with the reliable corpus *TC1*.

**Table 3: DRS Performance**

DRS Train Corpus	Best $T_d$ (%)	wDia Size ( $10^6$ words)	WER (%)	ChER (%)
TC0	-	-	1.13	0.256
TC1	-	-	0.59	0.133
TC2	11	25.9	0.94	0.211
TC3	13	37.0	0.92	0.207
TC4	12	61.4	0.79	0.177
TC5	10	80.9	0.55	0.124
TC6	13	91.7	0.56	0.126
TC7	11	117.3	0.53	0.120
TC8	12	179.5	0.52	0.116

**Table 4: ASR Language Models**

ASR Language Model	ASR Train Corpus	DRS Train Corpus
LM0	TC2+TC3+TC4	-
LM1	TC2+TC3+TC4	TC1
LM2	TC2+TC3+TC4 (our methodology)	TC2+TC3+TC4 (our methodology)
LM3	[9]	TC0
LM4	[9]+Meetings	TC1
LM5	[9]+TC8 (our methodology)	TC8 (our methodology)

**Table 5: ASR Output Quality**

ASR Language Model	Perplexity	OOV Rate (%)
LM0	154.9	2.49
LM1	150.6	2.44
LM2	148.2	2.31
LM3	176.9	3.13
LM4	151.2	2.95
LM5	129.0	2.35

## 4.2. ASR Experimental Set-up and Results

In order to measure the effect of the DRS on ASR output performance, six ASR language models are constructed, and presented in Table 4. *LM0-LM2* utilize as train corpus a combination of all news corpora acquired online. For *LM0*, no diacritics restoration is performed. This language model is designed to evaluate the usage of raw unreliable text for ASR training.

*LM1* makes use of a DRS constructed from *TC1*, a reliable text corpus, and is designed to evaluate the “traditional” methodology of using a large reliable corpus to train a DRS to correct unreliable online text before ASR training. *LM2* makes use of a DRS constructed with our methodology, from unreliable online text resources, and is designed to illustrate the case where no reliable resources exist (as is also the case with *LM0*). For *LM2*, our methodology is employed to generate a DRS and restore diacritics to the ASR train corpus. Because the training corpus is identical, and only diacritics restoration differs, *LM0-LM2* are directly comparable.

In addition, we analyze three more language models, *LM3-LM5*, which are intended to provide a comparison to previous work and more realistic use-case scenarios for our methodology. *LM3* is built from the training corpus presented in [9], which was corrected with the DRS presented in [8]. This language model will serve as a performance benchmark.

*LM4* is constructed from the ASR training corpus from [9], to which we added the *Meetings* corpus. For diacritics restoration, we utilize a DRS trained with *TC1*, the largest reliable text corpus available. *LM4* will serve as an indicator of what performance can be achieved by updating previous work with reliable text which has become available since [8] and [9] were published.

Finally, *LM5* is trained utilizing our methodology and all text resources currently available, including the reliable resources from previous work, the *Meetings* corpus, and the unreliable news text (*Antena3*, *Realitatea* and *Libertatea* corpora) acquired online. *LM5* serves as an indicator of the performance of a real-world ASR system trained with our methodology. In order to be able to directly compare previous work to our methodology, all ASR language models are evaluated on reference recordings and transcriptions previously utilized in [9].

With regard to the improvement in ASR output quality, evaluation results are presented in Table 5. *LM1* and *LM2* have roughly equivalent performance, which demonstrates the viability of our methodology, since *LM2* is created utilizing no reliable text at all. Both language models perform slightly better than *LM0*, which does not utilize diacritics restoration.

Regarding the more realistic ASR language models, *LM3-LM5*, we can observe a decrease in perplexity and OOV rate, compared to *LM3*, with the addition of the *Meetings* training corpus in *LM4*, and utilizing the DRS

trained with *TC1*. Furthermore, with *LM5* (generated from *TC8* with  $T_d$  at 12), perplexity and OOV rate are reduced by a 14% and 20% respectively compared to *LM4*.

## 5. CONCLUSION AND FUTURE WORK

We have presented a methodology which generates a diacritics restoration system from an unreliable corpus of raw text data, acquired from the web. Through a process of filtering, the raw text corpus is divided into a trusted sub-corpus and an untrusted sub-corpus, with respect to the use of diacritics. The trusted corpus is used to train a diacritics restoration system. Using this system, diacritics are restored to the untrusted sub-corpus, which is then used in conjunction with the trusted corpus to train a language model for automatic speech recognition. We have demonstrated experimentally that the methodology can be utilized to generate a diacritic restoration system of similar or better quality to those presented in previous work. This paper presents the best performing diacritics restoration system for the Romanian language, with regard to WER and ChER. Future work will focus on the development of an adaptive mechanism for automatically setting the diacritics separation threshold.

## ACKNOWLEDGEMENTS

This study has been partially developed with the financial support of the Romanian-American Foundation. The opinions, findings and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the Romanian-American Foundation.

## REFERENCES

- [1] D. Tufiş and A. Chiţu, "Automatic diacritics insertion in Romanian texts", in *Proceedings of the International Conference on Computational Lexicography*, Pecs, Hungary, pp. 185-194, 1999.
- [2] P. Brown, P. Desouza, R. Mercer, V. Pietra, and J. Lai, "Class-based n-gram models of natural language", in *Computational Linguistics*, vol. 18, no. 4, pp. 467-479, 1992.
- [3] I. Skadina, A. Vasiljevs, R. Skadins, R. Gaizauskas, D. Tufis, and T. Gornostay, "Analysis and evaluation of comparable corpora for under resourced areas of machine translation", in *The 5th Workshop on Building and Using Comparable Corpora*, pp. 17, 2012.
- [4] K. Scannell, "The Crubadan Project: Corpus building for under-resourced languages", in *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, vol. 4, pp. 5-15, 2007.
- [5] D. Tufiş and A. Ceaşu, "DIAC+: A professional diacritics recovering system", in *Proceedings of LREC*, 2008.
- [6] C. Ungurean, D. Burileanu, V. Popescu, C. Negrescu, and A. Dervis, "Automatic diacritic restoration for a TTS-based e-mail reader application", in *UPB Scientific Bulletin, Series C*, vol. 70, no. 4, pp. 3-12, 2008.
- [7] C. Ungurean and D. Burileanu, "An advanced NLP framework for highquality Text-to-Speech synthesis", in *The 6th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1-6, 2011.
- [8] H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, "ASR domain adaptation methods for low-resourced languages: Application to Romanian language", in *Proceedings of the 20th IEEE European Signal Processing Conference (EUSIPCO)*, pp. 1648-1652, 2012.
- [9] H. Cucu, A. Buzo, L. Besacier, and C. Burileanu, "SMT-based ASR Domain Adaptation Methods for Under-Resourced Languages: Application to Romanian", in *Speech Communication*, vol. 56, pp. 195-212, 2013.