On the Romanian rhyme detection

Alina Maria CIOBANU, Liviu P. DINU
University of Bucharest, Faculty of Mathematics and Computer Science,
Centre for Computational Linguistics, Bucharest, Romania
alinamaria.ciobanu@yahoo.com,ldinu@fmi.unibuc.ro

ABSTRACT

In this paper we focus on detecting Romanian words without rhymes, using knowledge about stressed vowels and syllabification. We also investigate quantitative aspects and the etymological origins of the Romanian words without rhymes.

KEYWORDS : Romanian language, syllables, rhyme

Proceedings of COLING 2012: Demonstration Papers, pages 87–94,
COLING 2012, Mumbai, December 2012.

1 Introduction

Rhyme represents the correspondence of final sounds of words, beginning with the rightmost stressed vowel. Hence, two words rhyme if their final stressed vowels and all following phonemes are identical (Reddy and Knight 2011).

In Romanian language the accent is variable and therefore cannot be determined in a deterministic manner. It can differentiate between words (“mozaic” – adjective, “mozaic” - noun) and grammatical forms (DOOM dictionary). Syllabification is a challenging and important task, considering that a rigorous research on syllable structure and characteristics cannot be achieved without a complete database of the syllables in a given language (Dinu and Dinu 2006, Dinu and Dinu 2009). Some attempts have been made for the automation of syllabification. Dinu and Dinu (2005) proposed a parallel manner of syllabification for Romanian words, using some parallel extensions of insertion grammars, and in (Dinu, 2003) is proposed a sequential manner of syllabification, based on a Marcus contextual grammar.

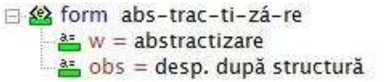
Identifying words without rhyme is an important problem for poets and especially for automatic or assisted poetry translation. Reddy and Knight (2011) emphasize another related research area – historical linguistics, as rhymes of words in poetry can provide valuable information about dialect and pronunciation at a given time. We propose an instrument which can provide rhyming words for a given input word and offers other features as well, such as detection of words without rhymes and clustering words based on syllable number. These tools contribute to determination of rhythm and metrics. This instrument can be very useful in identifying words with sparse rhyme. Reddy and Knight (2011) affirm that repetition of rhyming pairs is inevitable in collections of rhyming poetry and is partially caused by sparsity of rhymes. In addition, we focus on identifying the etymological origins for input words in a given language. This feature proves especially useful for words without rhyme.

2 Related work

To our knowledge, there are two sites that focus on finding rhymes for Romanian words (http://www.webdex.ro/online/dictionar/rime and http://www.spunetiparerea.ro/dictionar-de-rime/cauta-rime.php). Both of them accept an input word and identify rhyming words. However, both systems determine rather the longest common suffix than rhyme. To our knowledge, they do not provide the other features that we discussed: the ability to automatically identify all words without rhymes from the dictionary and clustering words based on syllable number.

3 On the rhyme detection

The dataset we used is a Romanian language resource containing 525528 words, including all inflectional forms (Barbu, 2008). For each word, the following pieces of information are provided: syllabification, type of syllabification (based on pronunciation or structure), position of the stressed vowels and inflectional form. Below is represented a word entry in our dataset. The “obs” field indicates that the word is syllabified based on its structure.



form abs-trac-ti-zá-re
w = abstractizare
obs = desp. după structură

FIGURE 1 – word entry in database

In order to determine the words that do not rhyme with other words, we focused on clustering them based on their rhymes.

1. For polysyllabic words, which had stressed vowels marked, we considered the substring beginning with the rightmost stressed vowel.
2. For monosyllabic words, which did not have stressed vowels marked, we had to focus on the position of the vowels within diphthongs and triphthongs in order to correctly determine the rhymes.
   - 2.1. Therefore, for ascendant diphthongs we considered the substring beginning with the second character of the structure and for descendant diphthongs we considered the substring beginning with the first character of the structure.
   - 2.2. We applied similar rules for triphthongs, considering the substring beginning with the third character of the structure for ascendant triphthongs and the substring beginning with the second character of the structure for balanced triphthongs.
3. Once all the words in the dataset were clustered based on their rhymes, we easily identified those that did not have any correspondent words and were, hence, without rhyme.

We identified 8851 different rhymes, among which the most frequent is "áţi", having 8142 corresponding words. 10808 words of our database (2.05%) do not have rhyme.

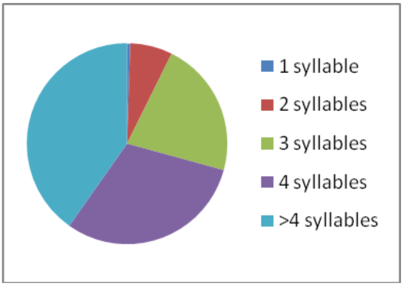


FIGURE 2 – Words with rhymes

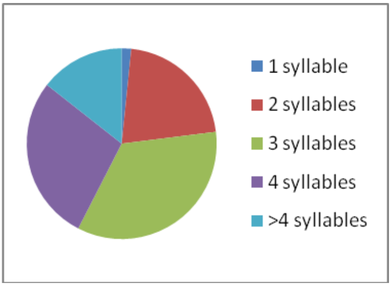


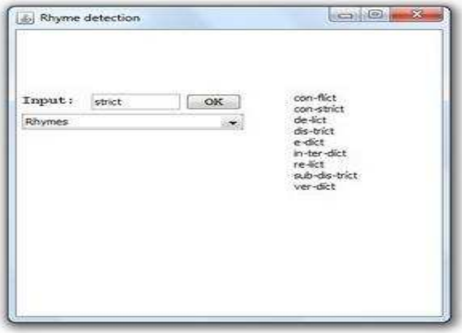
FIGURE 3 – Words without rhymes

Some observations can be derived from these charts. Most words with rhymes have more than four syllables, while most words without rhymes have three syllables. The number of monosyllabic words that do not have rhyme is small, only 174 out of 2894 do not rhyme with any other word. For each syllable cluster, the number of rhyming words is greater than the number of words without rhymes. The percentages in Table 1 do not add up to 100% because of the rounding.

	Dictionary		Words with rhymes		Words without rhymes	
1 syllable	2894	0.55%	2720	0.51%	174	0.03%
2 syllables	37235	7.08%	34923	6.64%	2312	0.43%
3 syllables	116806	22.22%	113073	21.51%	3733	0.71%
4 syllables	159911	30.42%	156877	29.85%	3034	0.57%
>4 syllables	208682	39.70%	207127	39.41%	1555	0.29%

TABLE 1– words distributed by number of syllables

Once we detected which words of the dataset rhyme with other words, we could implement the other features stated above. Our tool provides rhyming words for a given input, relevant in automatic or assisted poetry translation. Below are selected words that rhyme with the monosyllabic word “strict”, whose rhyme is “íct”. The output words are syllabified and their stressed vowels are marked. It can be easily observed that all retrieved words have the same rhyme as the input word.



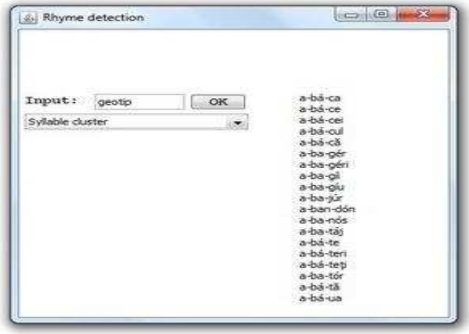
Input: strict

Rhymes

```
con-flict
con-strict
de-lict
dis-trict
e-dict
in-ter-dict
re-lict
sub-dis-trict
ver-dict
```

FIGURE 4 – Detecting word rhymes

Another important feature of our tool is the capability of clustering words based on their number of syllables. This feature proves very useful in identifying rhythm and metrics for poetry. In Figure 5 our tool identified words having an equal number of syllables with the Romanian word “geotip”, which is formed by three syllables (“ge-o-típ”).



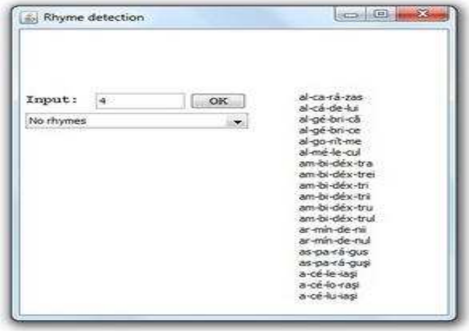
Input: geotip

Syllable cluster

```
a-bá-ca
a-bá-ce
a-bá-cei
a-bá-cul
a-bá-că
a-ba-gér
a-ba-géri
a-ba-gii
a-ba-giu
a-ba-júr
a-ban-dón
a-ba-nós
a-ba-táj
a-bá-te
a-bá-teri
a-bá-teţi
a-ba-tór
a-bá-tă
a-bá-ua
```

FIGURE 5 – Clustering words based on number of syllables

Our tool is able to identify words without rhymes, based on the algorithm we described above. In addition, the number of syllables can be selected, in order to retrieve only words without rhyme having the desired number of syllables. This feature is demonstrated below by selecting words without rhymes formed by four syllables.



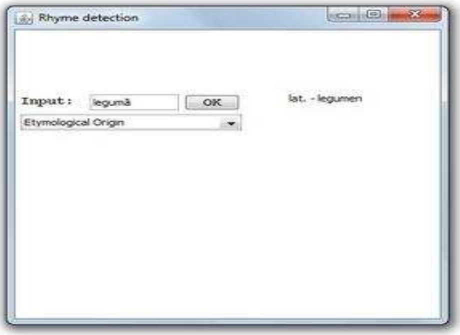
Input: 4

No rhymes

```
al-ca-rá-zas
al-cá-de-lui
al-gé-bri-că
al-gé-bri-ce
al-go-rít-me
al-mé-le-cul
am-bi-déx-tra
am-bi-déx-trei
am-bi-déx-tri
am-bi-déx-trii
am-bi-déx-tru
am-bi-déx-trul
ar-min-de-nii
ar-min-de-nul
as-pa-rá-gus
as-pa-rá-guşi
a-cé-le-iaşi
a-cé-io-raşi
a-cé-lu-iaşi
```

FIGURE 6 – Retrieving words without rhyme

After identifying words without rhymes, we were interested in establishing their etymological origin. In order to automatically detect this information, we used the dexonline.ro database (Copyright (C) 2004-2012 DEX online (http://dexonline.ro)), which gathers information from numerous Romanian dictionaries. We were thus able to determine the origins of the words in our dataset. This feature is valuable especially for words without rhymes. In Figure 7 our tool detected the origin of the word “legumă” and provided the Latin corresponding word (“legumen”).



Input: legumă

Etymological Origin

lat. - legumen

FIGURE 7 – Determining word’s etymological origin

Conclusion and perspectives

In this paper we present a method of identifying words without rhyme using knowledge regarding stressed vowels and syllabification applied on a dataset containing 525528 Romanian words, including all inflectional forms. Our main objective was to determine the words that do not have rhyme and to analyse their syllabic structure and etymological origin. The main result is that 2.05% of the words do not have rhyme. Most words with rhymes have more than four syllables, while most words without rhymes have three syllables.

Further, we presented a tool that is able to identify words without rhymes, to cluster words based on syllable number, to retrieve rhyming words for a given input and to identify etymological origins of the words. Being able to achieve all these pieces of information, our tool can be relevant in detection of words with sparse rhyme and automatic or assisted poetry translation. This research area has received attention in the past, but less effort has been spent on poetry analysis and translation until now (Greene et al 2010).

Acknowledgments

The research of Liviu P. Dinu was supported by the CNCS, IDEI - PCE project 311/2011, ”The Structure and Interpretation of the Romanian Nominal Phrase in Discourse Representation Theory: the Determiners.” We want to thank to Solomon Marcus for the helpful comments. Note that the contribution of the authors to this paper is equal.

References

Barbu, A.M. (2008): Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries, LREC 2008, May, 28-30, Marrakech, Marocco, 2008.

Dexonline.ro database - Copyright (C) 2004-2012 DEX online (http://dexonline.ro)

Dictionarul ortografic, ortoepic si morfologic al limbii romane (DOOM), Ed. Univers Enciclopedic Gold, Bucuresti, 2010.

Dinu, L.P. (2003). An approach to syllables via some extensions of Marcus contextual grammars. Grammars, 6(1), 1-12.

Dinu A. and Dinu L. P. (2005): A Parallel Approach to Syllabification, CICLing 2005, LNCS 3406, 83-87, 2005.

Dinu, A. and Dinu, L.P. (2006). On the data base of Romanian syllables and some of its quantitative and cryptographic aspects. In LREC 2006, May, 24-26, Genoa, Italy.

Dinu A. and Dinu L. P. (2009): On the behavior of Romanian syllables related to minimum effort laws. In C. Vertan, S. Piperidis, E. Paskaleva, M.Slavcheva (eds.): Proc. of Int. workshop on Multilingual resources, technologies and evaluation for central and Eastern European languages (workshop at RANLP 2009 conference), pp. 9-13, 14-16 September, 2009.

Greene E., Bodrumlu T. and Knight K. (2010): Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. EMNLP 2010: 524-533.

Reddy S. and Knight K. (2011): Unsupervised Discovery of Rhyme Schemes. ACL (Short Papers) 2011: 77-82.