Computer Science Journal of Moldova, vol.4, no.1(10), 1996

# An approach to the word division into syllables for Romanian language *

V.Demidov T.Verlan

**Abstract**

The solution of a problem of correct hyphenation of Romanian language words, and in general, problem of division them into syllables, becomes especially urgent today in view of absence of algorithm of word division into syllables in many widely used automated text processing systems exactly for Romanian language. This problem is nontrivial, whereas at a stage of word processing the phonetic information is inaccessible, although just this information is essential.

## Introduction

In laboratory “Theory and practice of programming” of the Institute of Mathematics of Academy of Sciences of Moldova the work on the Romanian language spelling checker creation and improvement is carried out during three years. Within the framework of this work a database of Romanian words was created. It continues to be filled up and at present totals 700000 words.

Alongside with other opportunities, the spelling checker contains the programs of the automatic inflexion formation for all parts of speech of Romanian language. Besides this a prompt system in a dialogue

*This work was partially supported by the Soros Foundation of Moldova, project No. 1 ED OPEN 96

mode works which permits to find faulty words in the text and offers variants of their replacement by a correct ones.

As we are also the active users of our spelling checker, it is clear for us the solution of which problems the latter could take on itself. One of these has become a problem of correct hyphenation of Romanian language words, that in general results in the problem of word division into syllables. This problem becomes urgent in view of lack of such algorithm just for Romanian language in many widely used systems of automated text processing.

## 1 Classification of the rules of word division into syllables

We shall present the classical rules of word division into syllables [1], which are based on letters' phonetic significance. The classical rules base on a vowel sequence - simple (**V**) and complex ( **Sv** - semivowels), and consonants (**C**) - also simple and complex, located between two vowels **V**.

### Vowels (V), simple or repeated

1. When there are two consecutive vowels **V**, a general rule of division into syllables is the following: **V-V** (*a-ud, mu-ze-e*).

   Note, that the division does not occur in the case of two repeated vowels, which form one sound, as a rule, in words of foreign origin (*week-end*).

### Complex vowels V and Sv

1. **VSvC** = **VSv-C**; *doi-nă.*
2. **VSvV** = **V-SvV**; *ba-ie, su-iai, cre-ia-ză.*

3. **VSvSvV = V-SvSvV**; *cre-ioa-ne.*

**Consonants C, simple, complex and repeated**

1. **VCV = V-CV**; *ma-să, re-ce, fa-ce, ve-che.* ( It is necessary to note, that the combinations "ph", "ch", "gh", "ck" - are the complex letters, forming a single sound: *ro-cker,ghi-o-cel, amphi-te-a-tru,che-ma-re*).

2. **VCC... V = VC-C... V**: *ar-tă,tor-rul,for-ţă.*

Below we describe exceptions of this rule.

a. In the case **VCCV** the rule **VCCV = V-CCV** is applied for the following combinations of consonants: "bl", "br", "cl", "cr", "fl", "fr", "hl", "pl", "pr", "tl", "tr", "vl", "vr" (*ca-blu,a-brupt, a-cla-ma, a-cru, A-fri-ca, si-hlă, cu-pru,a-tlet, pa-tru,e-vla-vi-e, co-vrig*).

   Note, that except the combinations listed above we have found one more, namely "gr", which was not described in the literature (*a-gre-ment, pro-gra-ma-re*).

b. In the case **VCCCV** the rule **VCCCV = VCC-CV** is applied for the following combinations of consonants: "lpt", "mpt", "mpţ", "ncş", "nct", "ncţ", "ndv", "rct", "rtf", "stm" (*sculp-tor, somp-tu-os, punc-tul, sfinc-şii, func-ţie,arc-tic, jert-fă*).

c. In the case **VCCCV** the rule **VCCCCV = VCC-CCV** is applied for the following combinations of consonants: "ngst", "rnbl", (*tung-stem, horn-blen-dă*).

d. In the case **VCCCV** for a combination of consonants "ngstr" the rule is the following: **VCCCCCV = VCC-CCCV** (*ang-strom*).

When developing our algorithm the rules described above were maximally taken into consideration. However, the specific character of

Romanian language does not permit completely formalize them. The vowels present main difficulty for division into syllables in Romanian language. They can be simple and complex (so called semivowels), stressed and unstressed, and the division rule depends on that, to which category the given vowel belongs (*frea-măt, re-cre-at, ghiont, ghi-o-cel*, etc). Besides this the ambiguity arises because of the mode in which different vowels combinations are perceived by ear (*au-gust, a-ur; leu-că, ne-u-tron*). When the word is entered, all that we can to find out about it is the consequence of vowels and consonants (**V** and **C**). The phonetic information is not accessible for us. Therefore, we can not implement the above described rules in all their completeness. These problems are well known and there are some attempts to describe them in literature ([2, 3]). However many situations are quite solvable, but rather by an artificial way. We shall describe below, how we managed to avoid some of the complexities.

## 2 Problem of diphthongs and threephthongs

Problem of diphthongs and threephthongs is the most difficult one in the process of word division into syllables for Romanian language. This complexity consists in following: the combinations of some vowels ("au", "ea", "ia", "ie", "io", "ii", "oa", "ua") can either form the diphthongs (*sea-ră, ier-ta, iod, au-gust*) or not (*re-al, scri-e-re, pi-on, ta-ur*) ([4]). The similar problems arise in the case of threephthongs too, for example: *chioa-ră, al-bi-oa-ra; a-pă-ra-iei, teh-no-lo-gi-ei.*

### Diphthong "ea"

We'll show by an example of combination "ea" how we managed to solve partially this problem. Having carried out the statistical analysis on the database, we have noticed, that it is possible to find some laws for the combinations "eal", "eat", "nea" and "rea". We have written the auxiliary programs of selection from the database the words, containing these combinations. So we found out the following:

**For "eat"** the words, containing combinations "de-at", "se-at", "ce-at", "me-at", "te-at", "re-at", "le-at", "ne-at", "be-at", are divided practically all, except several, which we mark specially in the program (*deseatină, dumneata, pleat, pomneata, veleată, şireată, tureatcă, trimeată*). The combination "beat" at the beginning of a word is not divided in general (*bea, beatnic, beato*), except the words, containing "be-a-tri", "be-a-ti". Note, that to capture all inflexions, our algorithm works with constant parts of words. For example, for processing the words *şireată, şireata, şireato*, we should put down into algorithm only constant part *şireat.*

**For "eal"** the following combinations: "neal", "osteal", "deal", "real" are subject to the certain laws:

- "Ne-al" at the beginning of word is always divided (*ne-al-tul, ne-a-li-ne-at*);
- "De-al" is usually not divided (*deal, is-co-dea-lă*), except the words, containing: "i-de-al", "de-alt-fel";
- "Oste-al" is divided;
- "Re-al" is an especially interesting case. There are some combinations of letters, after which "real" is always divided. We found out the following: "a", "anti", "aste", "bo", "co", "ce", "supra", "neo", "ne", "flo", "i", and "p". The examples of such words are: *a-re-al, an-ti-re-a-lism, as-te-re-al, co-re-a-li-ta-te, ce-re-a-le, flo-re-al, pre-a-la-bil.*

**For "nea"** the words, beginning with this combination of letters, usually are divided (*ne-a-pă-rat,ne-a-tent,ne-as-cul-ta-re*, etc), except the words, beginning with the following combinations: "neag", "neam", "neant", "neao" (*nea-gă, neamţ, neant*).

**For "rea"** the words, beginning with this combination of letters, usually are divided (*re-a-bi-li-ta, re-ac-tiv, re-a-dor-mi, re-a-ni-ma-re, re-a-u-di-e-re*, etc), except of the words *rea-văn, rea-zem.*

### Diphthongs and combinations “ie”, “ia”, “io”, “iu”, “ui”, “au”, “oa”

We also managed to solve partially the problem of these combinations. We have noticed (and the researches on the database with the help of programs of word selection with certain combinations have confirmed our supposition), that:

- If there are the consonants “g”, “c” and “ţ” before “ie” and “ia”, as well as if there is the consonant “ţ” before “io” , then there is a hyphen between vowels: *in-for-ma-ti-ci-an, i-ni-ţi-a-ti-vă, teh-no-lo-gi-a, ma-te-ma-ti-ci-e-ni-lor, nos-tal-gi-e, a-pa-ri-ţi-ei, so-lu-ţi-o-na-re.* The exceptions make the words *ciad, cian, cier, cia-un, cia-con, gia-că, giar-di-a.*

- If the word finishes with “iei” or “ier”, then the hyphen is the following: “i-e” (*a-ca-de-mi-ei,in-for-ma-ţi-ei,a-ni-ma-li-er*). Note, that for “iei” all the above said is true in the case, when there is a consonant before this combination (compare - *că-soa-iei* and *Ma-ri-ei*). The exceptions are: *miei, piei.*

  For the combination “ier” the exception makes the word *fier.*

- The combination “io” at the beginning, if it is followed by the consonant “n”, is usually divided, except the word “io-nat” (*i-o-ni-za-re, i-o-nic, i-o-no-sfe-ră*).

- The combinations “ui” and “iu” at the beginning and “iu” at the end are usually not divided: *ui-ta, ui-mi-re, iu-bi-re, a-u-riu.* The exceptions make the words *i-ui, u-ie-maş.*

- The combination “au” at the beginning of the word is usually divided (*a-u-di-a, a-ur, a-u-to-buz, a-u-to-crat*, etc.), except the words, containing the following combinations: “aug”, “au-rig”, “aus-cult”, “aus-tru”, “aut”, “au-to-daf”, “au-tum-nal”, “au-tu-ni”, “au-ver-si”.

- It is necessary to note especially the case, when the words are finished with the vowel “i”. For example, considering the combination of vowels and consonants, the words “arici” and “alipi” look the same: **VCVCV**. However, they are divided differently: *a-rici, a-li-pi.* Since the phonetic information is not accessible to us, so as in the previous case, we have rejected the last hyphen (*a-lipi, a-min-tiri, vîn-turi, fugi, duci, co-pi-ii, lă-măi*).

- In the words, finishing with vowel “u”, the hyphen is not allowed before this vowel, as far as it is the semivowel (*rău, me-diu, mereu*). The exception makes the word *con-ti-nu-u.*

## 3 Prefixes

- When there is one of the combinations “an”, “in”, “în” at the beginning of the word, and it is followed by the vowel, then an ambiguity appears . We shall explicate this by the following examples:

  a. In such a words, as *an-io-nit, an-o-re-xie, an-or-ga-nic, in-e-vi-ta-bil, in-ex-plo-ra-bil, în-ăs-pri, în-al-bit* - the combinations indicated above serve as the prefixes, and the hyphen should be put after them;

  b. In such a words, as *a-nor-mal, a-ni-mal, i-ni-ţi-al, i-no-va-ţie, î-năl-ţi-me, î-nă-un-tru* - according to the classical rules, the hyphen should be put after the vowel.

  To avoid the ambiguity, and in the first place taking into consideration the problem of the word division from line to line, when one letter is not left on a line, we have decided to reject the first hyphen in such cases: *anor-ga-nic, ine-vi-ta-bil, înăl-bit, ani-mal, ini-ţi-al, înăl-ţi-me.*

- In the combinations “nest”, “neşt”, “nesp” at the beginning of the word, the hyphen is put after “n” (*ne-sta-bil, ne-spus, ne-şti-ut, ne-spus*);
- The combination “post” at the beginning of a word is usually divided (*po-stal, po-sta, po-stă-var*), except of the words, marked in our program (*post-be-lic, post-ver-bal* and other).
- The combination “dez” at the beginning of a word is not divided (*dez-ac-ti-va, dez-ba-te, dez-vol-ta-re*), except the words “de-zast” and “de-zert”.
- In the words, beginning with prefixes “micro-”, “tele-” and “stereo”, which are followed by two consonants, the hyphen is put before these consonants: *mi-cro-sco-pi-e, te-le-graf, ste-re-o-chi-mi-e*;
- In the words, beginning with prefix “trans”, the hyphen is put, as a rule, after this prefix: *trans-for-ma, trans-a-mi-na-re, trans-hu-mant*, etc. The exception is made by the words, containing the following constant parts: “tran-să”, “tran-scri”, “tran-sept”, “tran-silv”, “tran-spir”, “tran-sis-tor”.

## 4 Compound words

The compound words by the mode of spelling are divided in the two groups:

- The words, which are always written with hyphen: *baba-oarba, bună-credinţă, prim-ministru, anglo-franco-italian*;
- The words, consisting of several words, which are written as a single word: *feldmareşal, concertmaistru, binefacere, botgros*;

We can not correctly divide all the words of the second group yet, for the lack of the necessary information when the word is entered: a) it is not known whether the word is compound; b) it is not known, of which words it consists. For example, the words *bi-ne-fa-ce-re, bu-nă-vo-in-ţă* will be divided by our algorithm correctly. But such compound words, as *feldmareşal, concertmaistru*, will be divided incorrectly. Each word of the first group is processed as a group of independent words, between which the hyphen is already known (*ba-ba-oar-ba, bu-nă-cre-din-ţă, prim-mi-nis-tru, an-glo-fran-co-i-ta-li-an*).

## 5 Conclusion

Thus, the algorithm of division into syllables of rather extensive class of Romanian language words is obtained. Certainly, we have not pretensions to the completeness. However, the testing showed, that 70 % of words in the texts from the scientific and art literature are divided correctly. It is essential. We should especially note that the Dictionary [5] was very useful for us because it contains the morphological information about words.

The algorithm can be developed more over, but to include some additions the further analysis of the database is necessary. It is the routine work, which takes a lot of time though. However, the main and most frequently met letter combinations our algorithm processes correctly. Therefore, it is effective enough.

## References

[1] Flora Şuteu, Elizabeta Şoşa, “Dicţionar Ortografic al Limbii Române”, (rom.), ATOS, Bucureşti, 1994.

[2] Felicia Şerban, Luciana Peef, Lidia Bibolar, Dana Bucerzan, “Baza de Date a Limbii Române Fonetică şi Fonologie”, (rom), Limbaj şi tehnologie, Ed. Academiei Române, Bucureşti, 1996.

[3] Amalia Todiraşcu, "Un Model bazat pe unificare pentru generarea vorbirii", (rom), Limbaj şi tehnologie, Ed. Academiei Române, Bucureşti, 1996.

[4] "Norme Ortografice, Ortoepice şi de Punctuaţie ale Limbii Române", (rom), Lumina, Chişinău, 1991.

[5] "Dicţionarul Ortografic, Ortoepic şi Morfologic al Limbii Române", (rom), ed. Academiei, Bucureşti, 1989.

V.Demidov, T.Verlan,
Institute of Mathematics,
Academy of Sciences of Moldova,
5 Academiei str., Kishinev,
2028, Moldova
phone: (373–2) 738073
e–mail: 22valea@math.moldova.su
21tverla@math.moldova.su

Received 3 December, 1996