On the data base of Romanian syllables and some of its quantitative and cryptographic aspects

Liviu P. Dinu[*], **Anca Dinu** [†]

[*]University of Bucharest, Faculty of Mathematics and Computer Science
Academiei 14, RO-70109, Bucharest, Romania
ldinu@funinf.cs.unibuc.ro
[†] University of Bucharest, Faculty of Foreign Languages,
Edgar Quinet 17, Bucharest, Romania
anca_d_dinu@yahoo.com

Abstract

In this paper we argue for the need to construct a data base of Romanian syllables. We explain the reasons for our choice of the DOOM corpus which we have used. We describe the way syllabification was performed and explain how we have constructed the data base. The main quantitative aspects which we have extracted from our research are presented. We also computed the entropy of the syllables and the entropy of the syllables w.r.t. the consonant-vowel structure. The results are compared with results of similar researches realized for different languages.

1. Introduction

In the last decade, building language resources and theirs relevance to practically all fields of Information Society Technologies has been widely recognized. The term language resources (LR) refers to sets of language data and descriptions in machine readable form, such as written or spoken corpora and lexicon, annotated or not, multimodal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, etc. The relevance of the evaluation for language technologies development is increasingly recognized. On the other hand, the lack of these resources for a given language makes the computational analyzes of that language almost impossible. The lexical resources contain a lot of data bases of linguistic resources like tree banks, morphemes, dictionaries, annotated corpora, etc. In the last years, one of the linguistic structures which regained the attention of the scientific community from Natural Language Processing area was the syllable (Kaplan and Kay 1994, Levelt and Indefrey 2001, Müller 2002, Dinu 2003, Dinu and Dinu 2005). New and exciting researches regarding the formal, quantitative, or cognitive aspects of syllables arise, and new applications of syllables in various fields are proposed: speech recognition, automatical transcription of spoken language into written language, or language acquisition are just few of them. On the other hand, an alterable word can be more easily reconstructed if we know its syllabification, fact that has important implications in cryptography.

One of the first lexical resources regarding syllables was the data base of Dutch syllables (Schiller et al., 1996). In the next section we will detail the reasons for our decision to build a data base of syllables for Romanian language.

The rest of the paper is structured as it follows. In section 2. we argue in favor of the existence of a syllable data base for the Romanian language. Section 3. comprises the way such a data base was build and we present some quantitative aspects emerged from the analyzes of the Romanian syllables. The final section is reserved to the conclusions and future works.

2. Data base designing

A rigorous study of the structure and characteristics of the syllable is almost impossible without the help provided by a complete data base of the syllables in a given language. A syllable data base has not onely a passive role of description, but an active role in application as speech recognition. Also, the psycho-linguistic investigation could greatly benefit from the existence of such a data base (Menzerath, 1954, Levelt and Indefrey, 2001). These are some of the reasons which provided our motivation for creating a syllable data base for the Romanian language. Here are two of the main problems we were confronted to when building the data base:

1. How to choose the corpus in order to obtain a representative syllable data base for the Romanian language?
2. Once we get such a corpus, its dimensions demanded an algorithm for automate syllabification, given that it would be impossible otherwise to manually syllabify it.

In order to overcome the first problem, we used as a corpus the DOOM dictionary (1982). However, this solution is far from being perfect: even though this choice guarantees for the presence of all Romanian syllables as *types*,we do not get any information regarding the number of syllables as *tokens*. In other words, the frequency factor related to its usage is disregarded, each word from the dictionary being syllabified onely once. This is not in accordance to the fact that words have different occurrence frequency in the spoken (or written) language, frequency given by their capacity to form locutions, their polysemy, etc. (see the criteria for building the main lexical vocabulary, M. Dinu, 1996). The fact that for Romanian language an unanimously accepted and representative corpus (containing belletristics, scientific papers, drama, journallism, etc) does not exist, the need for an exhaustive data base for the Romanian syllables and the existence of DOOM in electronic format were sufficient reasons to choose the DOOM as the corpus to use. In some future work we hope to be able to present results obtained by analyzing a corpus which meets all the upper requirements and compare them to the results in this paper.

Regarding the second problem, the main obstacle was to extract the rules of syllabification and to adapt them to the computer requirements, without knowing the word accent. To solve this problem, we divided the rules of syllabification into two classes. The first one is formed by the rules which apply to a consonantal sequence of one up to five (the maximum length of a consonantal sequence in Romanian language). We formalized this rules completely for the computer requirements, thus the algorithm we proposed correctly syllabifies any consonantal sequence. The second class is formed by the rules which apply to a sequence of vowels of two up to six (the maximum length of a sequence of vowels in Romanian language). We observed that a sequence of vowels has regular behavior regarding its syllabification depending on the sequences of letters that succeeds and precedes it. Based on this observation, we proposed a set of rules of syllabification for sequences of vowels and we formalized them. The algorithm based on these rules (both for sequences of consonants and sequences of vowels) is described in (Dinu, 1997). This algorithm does not syllabify correctly 100%, thus some of the obtained syllables could be *false* syllables, perturbing the frequency of syllables. However, these perturbations are acceptable, not significatively influencing the data base we have constructed.

3. The analysis of the Romanian syllables

In this section we present the main results of our research. Before that, we shortly present a pioneering work done by Roceric (1968) in the sixties.

3.1. Previous works

Alexandra Roceric Alexandrescu presented in 1968 a quantitative study of the phonological structure, for the Romanian language. A. Roceric used belletristics and *Dicţionarul Limbii Române Moderne* by Cândrea.

The first part of this study is dedicated to some quantitative analyze of consonants and vowels. She observes that the ratio vowels-consonants is similar to the same ratio in other languages. She presents a series of combinatorial characteristics of phonemes, some distributional classes, phonemes frequencies, etc.

The second part of the study investigates the syllable and the word. The main consonant-vowel structures of Romanian words are determined. After dividing 3.700 words extracted from different texts, A. Roceric identifies 15 possible types which can appear inside a word in initial position, 10 possible types which can appear in median position and 17 which can appear in final position.

The author also treats the problem of the possible combination between syllables. Because the words formed by two syllables are most frequently used, she presents the combinatorial structure of these words, which are organized in 64 different types. Similar analyzes have been made for the words formed by 3, 4, 5 and 6 syllables. The most adherent syllabic structure is CV (consonant-vowel): for each number 2, 3, 4, 5 or 6 of syllables, the most frequent words have the structures: *CV-CV, CV-CV-CV, CV-CV-CV-CV, CV-CV-CV-CV-CV CV-CV-CV-CV-CV-CV*, respectively.

We must say that the Romanian linguistic school is represented since the XIX-th century by linguists as A. Cihac and B. P. Haşdeu who anticipated the use of statistic method in linguistics; in the IV-th decade of the XX-th century Pius Servien and Matila Ghyka (in collaboration with G. D. Birkhoff) are the first who introduced the mathematical models in poetics. A synthesis of Romanian research in mathematical and computational linguistics, having over 500 titles and over 120 Romanian authors is presented in (Marcus, 1978). The papers presented there are grouped in 7 categories: statistical linguistics, algebraic linguistics, analytical models, generative models, mathematical and computational poetics, computational linguistics, applications of mathematical linguistics in science and art.

3.2. Quantitative results based on DOOM corpus

The previous work was limited to a small corpus. Also, it is obviously that the lack of computational tools combined with the poor computational analysis of the syllabification were other impediments for Roceric's work.

In our research we used an approximatively exhaustive corpus, namely Orthographically, Orthoepic and Morphological Dictionary (DOOM) of Romanian language.

The corpus we used (DOOM) contains $N_{words} = 74.276$ words. We automatically syllabified the words using the algorithm presented in (Dinu, 1997) and we introduced the obtained syllables in a data base having the following fields: the syllable, its length, its vowel-consonant structure, the frequency of appearance of the syllable in a word on the first, median and last position, the frequency of appearance of the syllable as a single word, the total frequency (i.e. the sum of the upper frequencies), the combining possibilities of the syllable (i.e. which are the syllables which can follow it and can be followed by it).

The analysis of this data base allowed us to extract a series of quantitative and descriptive results for the syllables of Romanian language:

1. we identified $N_{Stype} = 6496$ (*type syllables*) in Romanian language. The total number of syllables (*token syllables*) is $N_{Stoken} = 273261$. So, the average length of a word measured in syllables is $Lwords_{syl} = N_{Stoken}/N_{words} = 273261/74276 = 3,678$.

2. The 74276 words are formed of $N_{letters} = 632702$ letters. So, the average length of a word measured in letters is $Lwords_{let} = N_{letters}/N_{words} = 632702/74276 = 8,518$.

3. In order to characterize the average length of a syllable measured in letters we investigated two cases:

   (a) the average length of the *token syllables* measured in letters is: $Lsyl_{token} = N_{letters}/N_{Stoken} = 632706/273261 = 2,315$

(b) The *type syllables* are formed of $N_{Tletters} = 24406$ letters. Thus, the average length of a *type syllable* measured in letters is $Lsyl_{type} = N_{Tletters}/N_{Stype} = 24406/6496 = 3,757$

4. The number of consonant-vowel structures which appear in the syllables is 56. Depending on the type-token rapport, the most frequent consonant-vowel structures are:

(a) for the *type syllables*: see Table 1.

C-V structure	Frequency	Percentage
cvc	1448	22%
ccvc	913	14%
cvcc	705	10%
cvcv	523	8%
cvvc	357	5%
ccv	354	5%
cvv	314	4%
cvccv	255	4%
ccvcc	223	3%
ccvv	166	3%
ccvcv	160	2%
cv	151	2%
ccvvc	92	1%
vc	89	1%
cccvc	76	1%
vcc	71	1%
ccvccv	66	1%
cccv	62	1%
vvc	59	1%
cvvcc	49	1%

Table 1: The statistics of type-syllables

(b) for the *token-syllables*: see Table 2.

C-V structure	frequency	percentage
cv	146744	53%
cvc	48139	17%
v	23707	8%
ccv	17418	6%
vc	11048	4%
cvv	6660	2%
cvcc	5684	2%

Table 2: The statistics of token-syllables

It is remarkable that the seven structures from Table 2 (i.e. 12% of the all 56 structures) cover approximatively 95% of the total number of the existent syllables.

5. the most frequent 50 syllables (i.e. 0,7% of the syllables number N_{Stype}) have 137662 occurrences, i.e. 50,03% of N_{Stoken}.

6. the most frequent 200 syllables cover 76% of N_{Stoken}, the most frequent 400 cover 85% of N_{Stoken} and the most frequent 500 syllables (i.e. 7,7 % of N_{Stype}) cover 87% of N_Stoken. Over this number, the percentage of covering rises slowly.

7. the first 1200 syllables in there frequency order cover 95% of N_{Stoken}.

8. 2651 syllables of N_{Stype} occur onely once (hapax legomena).

9. 5060 syllables (i.e. 78%) of N_{Stype} occur less then 10 times. These syllables represent 11960 syllables (4% of N_{Stoken}).

10. 158941 syllables (58% of N_{Stoken}) are formed of 2 letters; the syllables formed of 3 letters represent 27% of N_{Stoken}, those formed of 1 letter represent 9% of N_{Stoken} and those formed of 4 letters represent 6% of N_{Stoken}.

11. We computed the entropy of syllables, using the formula:

$$H_{syl} = -\Sigma_{i=1}^{6496} p_i log_2 p_i,$$

where p_i is the occurrence probability of the syllable situated on the i-th position in the classification obtained by ordering the syllables in decreasing order of their total frequencies. The probability p_i is computed as the ratio between the total frequency of the syllable situated on the i-th position and the total number of occurrences N_{Stoken}.

Thus, we obtained that the value of the syllable entropy is:

$$H_{syl} = 8,621$$

12. We also computed the entropy of syllable w.r.t. the C-V structures, using the formula:

$$H_{cv-syl} = -\Sigma_{i=1}^{56} p_i log_2 p_i,$$

where p_i is the occurrence probability of C-V structure of the syllable situated on the i-th position w.r.t. the order of occurrence frequency. We obtained that the value:

$$H_{cv-syl} = 2,30,$$

which is near to the values obtained by Edmond Nicolau (1962) or Alexandra Roceric (the value they obtained is 2,63).

4. Conclusions and future works

The linguists refused to accord to the syllable the status of structural unit of the language, as opposed to the units as the phoneme and the morpheme. As a consequence, the mathematical models of the syllable failed to equal the complexity of the morpheme and phoneme mathematical models. Opposite to the lack of qualitative insight regarding the syllable, the quantitative, statistic nature of the syllable was intensely studied.

In this paper we argued for the need to construct a database for Romanian syllable. We explained how we have constructed the data base, the corpus which we have used, the computational aspects of the syllabification and we presented some of the the main quantitative aspects which we have extracted from our research. Also, a series of results (like the adherence of syllables, the entropy calculus or the syllabification study) can be used in cryptography.

Some of the results presented in this paper are similar to other results reported for different languages (e.g. Schiller et. al. , 1996, for Dutch syllables.) and confirm a series of empiric laws. Other results of this research are reported for the first time here. In a future work we hope to be able to present results obtained by analyzing a corpus of spoken Romanian language other then the one we used (DOOM) and compare them to the results in this paper.

Acknowledgement 1 *The work of this paper was done with financial support from SCRIPT 8/2005 project from the SECURITATE Romanian Research Program. We would like to thank the referees for their comments which helped improve this paper. We would like also to thank to Emil Ionescu and Beatrice Popescu for theirs support during the designing of the syllables database.*

5. References

Dinu, L.P. The alphabet of syllables with applications in the study of rime frequency. *Analele Univ. Bucureşti*, XLVI-1997, 39-44, 1997.

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

A. Dinu, L.P. Dinu. A parallel approach to syllabification. In *A. Gelbukh (Ed.): CICLing 2005. LNCS 3406*, 83-87, 2005.

Dinu, M. *Personalitatea limbii române*. Ed. Cartea Românească, Bucureşti, 1996.

D.O.O.M. Dicţionarul ortografic, ortoepic şi morfologic al limbii române. Ed. Academiei, Bucureşti, 1982.

Kaplan, R.M. and M. Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3), 331-379, 1994

Levelt, W.J.M., P. Indefrey. The Speaking Mind/Brain: Where do spoken words come from. În *Image, Language, Brain*, eds. A. Marantz, Y. Miyashita, W. O'Neil, pp. 77-94. Cambridge, MA: MIT Press, 2001.

Marcus, S. Mathematical and computational linguistics and poetics. *Revue Roumain de Linguistique*, XXIII, 559-588, 1978.

Menzerath, P. Die Architektonik des deutschen Wortschatzes. În *Phonetische Studien*, Heft 3. Bon: Ferd. Dümmlers Verlag, 1954.

Müller, K. *Probabilistic Syllable Modeling Using Unsupervised and Supervised Learning Methods* PhD Thesis, Univ. of Stuttgart, Institute of Natural Language Processing, AIMS 2002, vol. 8, no.3, 2002

Nicolau, E. Langage et stratégie. *Cahiers de linguistique théorique et appliquée*, 1, 153-179, 1962.

Petrovici, E. Le pseudo *i* final du roumain. *Bulletin Linguistique*, 86-97, 1934.

Roceric-Alexandrescu, A. *Fonostatistica limbii române*. Ed. Academiei R.S.R., 1968.

Rosetti, A. *Introducere în fonetică*, Ed. Ştiinţifică, Bucureşti, 1963.

Schiller, N., A. Meyer, H. Baayen. A Comparision of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics*, 3, 1, 8-28, 1996.

Vasiliu, E. *Fonologia limbii române*. Ed. Ştiinţifică, Bucureşti, 1965.