On the behavior of Romanian syllables related to minimum effort laws

Anca Dinu
University of Bucharest
Faculty of Foreign Languages
and Literature
Edgar Quinet 5-7
Bucharest, Romania,
anca_d_dinu@yahoo.com

Liviu P. Dinu
University of Bucharest
Faculty of Mathematics
and Computer Science
Academiei 14, 010014
Bucharest, Romania,
ldinu@funinf.cs.unibuc.ro

Abstract

The main goal of this paper is to investigate the behavior of Romanian syllables related to some classical minimum effort laws: the laws of Chebanow, Menzerath and Fenk. The results are compared with results of similar researches realized for different languages.

Keywords

syllable, minimum effort laws

1 Introduction

In the last decade, the building of language resources (LR) and theirs relevance to practically all fields of Information Society Technologies has been widely recognized. LRs cover basic software tools for their acquisition, preparation, collection, management, customisation and use and are used in many types of applications (from language services to e-learning and linguistic studies, etc.) The relevance of the evaluation for language technologies development is increasingly recognised. 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 base of linguistics resources like tree banks, morphemes, dictionaries, annotated corpora, etc. In the last years, one of the linguistics 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 and Dinu 2005a,b). 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.

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 only a passive role of description, but an active role in applications as speech recognition. Also, the psycholinguistic investigation could greatly benefit from the existence of such a data base. These are some of the reasons which provided our motivation for investigating the behavior of Romanian syllable related to some cognitive laws, based on the corpus of Romanian syllable extracted from DOOM.

2 Motivation

The linguists refused to accord to the syllable the status of structural unity 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.

From the point of view of the language acquisition, the syllables are the first linguistical units learned during the acquisition process. Numerous studies showed that the children's first mental representation is syllabic in nature, the phonetic representation occurring only later.

Each language has its own way of grouping the sounds into syllables, as a result of its structure. The grouping of the syllables takes place depending on the innate psychic inclination of the group. If the vowels in a word are suppressed and only the consonants remain, the word form can be reconstructed with a high probability, when the syllabification of the word is known. This shows that from the existence of the consonant one can deduce the presence of the vowel, so one can determine the graphical form of the syllable an of the whole word. These aspects may have application in cryptography.

Numerous physiological experiments concerning the syllable are realized between the second part of the XIX-th century and the first part of the XX-th century. The experiments from 1899 made by Oussoff showed that the syllable does not always coincide with the respiratory act, because, during a single expiration, more then one syllable can be produced. In 1928 Stetson also showed that the syllable synchronizes with the movement of the thoracic muscles: each new movement of the muscles produces a new syllable (cf. Rosetti, 1963).

The psycholinguistic elements are situated inside the speech production area. Experiments revealed the presence of a library of articulatory pre-compiled

Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009, Bulgaria, pages 9–13

routines, which is accessed during the speech production process. In 1994 these observations leaded to the so-called *mental syllabary*. The theory of Levelt and Wheeldon (1994) assumes the existence of this *mental syllabary*: for frequently used syllables there is a library of articulatory routines that is accessed during the process of speech production. The adjoining of such syllabic gesture generates the spoken word and greatly reduce the computational cost of articulatory programs.

These aspects determined us to study and analyze the syllable. In the following we will focus on the lexical (not phonological) aspects of the syllable.

3 Quantitative aspects of the syllable

Opposite to the lack of qualitative insight regarding the syllable, the quantitative, statistic nature of the syllable was intensely studied.

Determining the optimal values of the length of sentences and of the words depending on the certain groups of readers may prove to be very useful in practical application. By optimum value we understand the value for which the level of comprehensibility is the biggest for the class of readers. Knowing this value should be especially important for the teachers and for publishers who print text books. The main conclusion of (Elts and Mikk, 1996) is that, for a good understanding of a text, the length of sentences in the text must be around the average length of sentences. Some optimum values are presented in the next table:

The optimal length of the words (Bamberge, Vanecek, 1984-cf. Elts and Mikk, 1996) (the first row is the readers' level, the second row is the length of words in syllables, and the third row is the length of words in letters):

Reader's level								
4	5	6	7	8	9	10	11	12
1.62	1.68	1.72	1.80	1.88	1.91	1.99	2.08	2.11
6.16	6.39	6.54	6.84	7.15	7.26	7.57	7.91	8.02

Another experiment on 98 students which were given 48 texts, produced the following optimal values:

	Level 8	Level 10
Optimal words (in letters)	8.53	8.67
Optimal sentences (in letters)	71.5	76.0

3.1 On the data base of Romanian syllables

In order to properly investigate the cognitive aspects of the syllables (often embedded in *minimum effort laws*), it is necessary to have a data base of syllables. In [9] a such database of Romanian syllables is presented. We list here some of the main results of this study, with possible cognitive implications. Based on this database, in the next section we will investigate the behavior of Romanian syllables related to some cognitive laws.

Based on the DOOM dictionary, which contains $N_{words} = 74.276$ words, the following series of quantitative and descriptive results for the syllables of Romanian language was extracted ([9]):

1. it was 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, two cases were investigated:

   (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*: *cvc* (22%), *ccvc* (14%), *cvcc* (10%).

   (b) for the *token-syllables*: *cv*(53%), *cvc* (17%), *v* (8%), *ccv* (6%), *vc* (4%), *cvv* (2%) and *cvcc* (2%).

   It is remarkable that these last 7 structures (i.e. 12% of the 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 only 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}.

The upper results are similar to other results, from different languages. For Dutch (cf. Schiller et al., 1996), the first 500 *type syllables*, ordered after their frequency,($\approx 5\%$ of the total number of *type syllables*), cover approximatively 85% of the total number of *token syllables*. For English, the result is similar, the first 500 syllables cover approximatively 80% of the total number of the *token syllables*. This results support the *mental syllabary* thesis.

4 The laws of Chebanow, Menzerath and Fenk for Romanian syllables

Several studies proposed laws of the *minimum effort type*: the famous Zipf's law, Menzerath's law which states that the bigger the number of syllables in a word, the lesser the number of phonemes composing these syllables. In cognitive economy terms, this means that *The more complex a linguistic construct, the smaller its constituents*. Fenk proposes another three forms of this law:

1. The bigger the length of a word, measured in phonemes, the lesser the length of its constituent syllables, measured in phonemes.

2. The bigger the average length of sentences, measured in syllables, the lesser the average length of syllables, measured in phonemes.

3. There is a negative correlation between the length of sentences, measured in words, and the length of the words, measured in syllables.

In this section we investigate the behavior of Romanian syllables related to the three above mentioned laws.

4.1 Chebanow's law

An intens studied problem in quantitative linguistics was the one regarding the existence of a correlation between the words' length (in syllables) and theirs occurrence's probability. In 1947, Chebanow investigated 127 Indo-European languages and he proposed a Poisson type law for the above problem.

For each particular language, he used a large number of texts to obtain the frequency of words. Denoting by $F(n)$ the frequency of a word having n syllables and by $i = \frac{\sum nF(n)}{\sum F(n)}$ the average length (measured in syllables) of the words, Chebanow proposed the following law between the average i and the probability of occurences $P(n)$ of the words having n syllables:

$$P(n) = \frac{(i-1)^{n-1}}{(n-1)!}e^{1-i}.$$

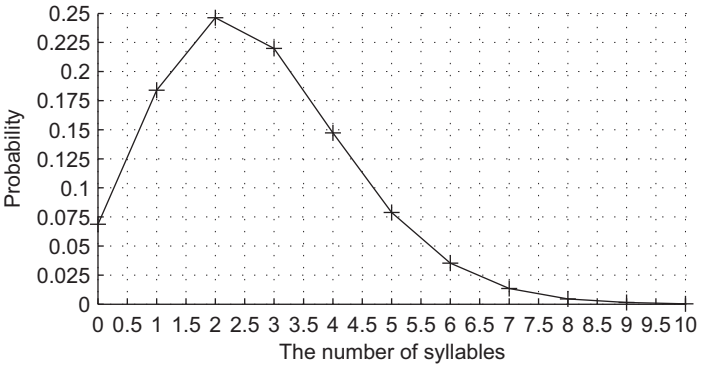

Fig. 1: *The Poisson distribution of length of words (parameter equal to 2.678)*

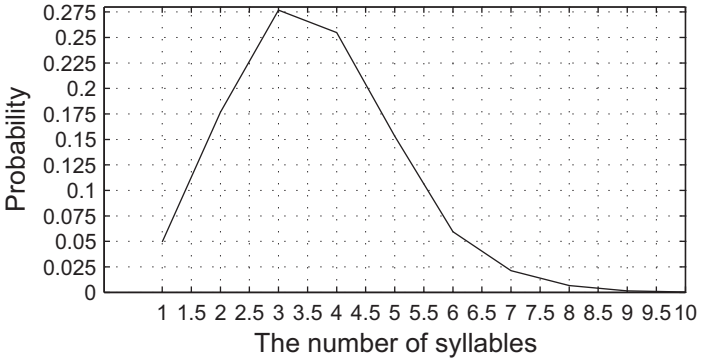

Fig. 2: *The probability distribution of the length of words*

We checked the Chebanow's law on the data base of Romanian syllables and we obtained a strong similarity between the Poisson's distribution (Fig.1) and the distribution of length (in syllables) of words (Fig. 2):

Remark 1 *It is important to see that the graphic from Fig. 2 must be translated with 1 to the left in order to overlap with Chebanow's law (probability $P(n)$ of the words of length n is the Poisson distribution with parameter $n-1$).*

Remark 2 *In the Fig. 1 we represented the following Poisson's distribution (the average length of word is 3.678, so we have to use the value 3.678-1=2.678, cf. Chebanow's law) :*

$$P(n) = \frac{2.678^n}{n!}e^{-2.678}.$$

4.1.1 Menzerath's law

We check the initial Menzerath's law, namely the one regarding a negative correlation between the length of a word in syllables and the lengths in phonemes of its constitutive syllables. The Fig. 3 shows that the law is satisfied.

4.1.2 Fenk's law

Fenk (1993) observed also that the bigger the length of a word, measured in phonemes, the lesser the length of

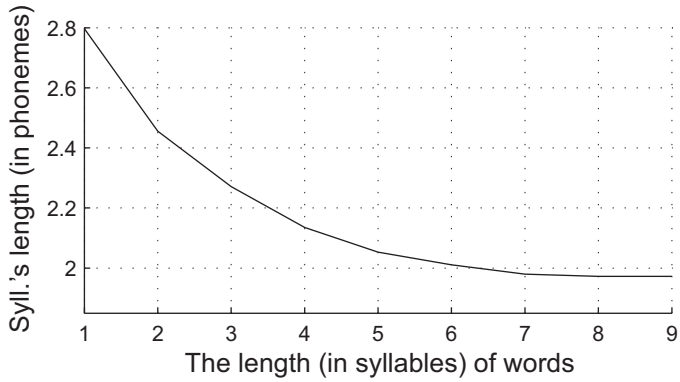

Fig. 3: *The Menzerath's law: The more syllables in a word, the smaller its syllables*

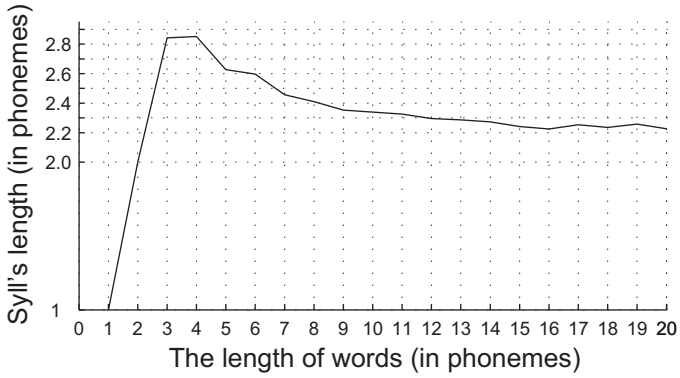

Fig. 4: *The Fenk's law: The more phonemes in a word, the lesser phonemes in its syllables*

its constituent syllables, measured in phonemes. We checked this correlation and the Fig. 4 confirms the first Fenk's law:

5 Conclusion and future works

In this paper we have presented some quantitative observations obtained from the analyze of a data base of Romanian syllables and we checked the behavior of the laws of Chebanow, Menzerath and Fenk for Romanian syllables. All of our results are similar to the results of other researches from different other natural languages (e.g. English, Dutch, Korean, cf. Schiller et. al 199 6, Choi 2000) . In some 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.

Acknowledgements: Research supported by CNCSIS, PN2-Idei, Project 228 and University of Bucharest, project IdeiUnibuc 17.

References

[1] Alekseev, P.M. Graphemic and syllabic length of words in text and vocabulary. *Journal of Quantitative Linguistics*, 5, 1-2, 5-12, 1998.

[2] Altmann, G. Prolegomena to Menzerath's law. În *Glottometrika 2*, 1-10, ed. R. Grotjahn, Bochum, 1980.

[3] Altmann, G. Science and linguistics. În *Contributions to quantitative linguistics*, eds. R. Köhler, B. B. Rieger. Kluwer Academic Publishers, Netherlands, 1993.

[4] Chebanow,S.G. On conformity of language structures within the Indoeuropean family to poisson's law. *Comptes rendus de l'Academie de science de l'URSS*. 55(1947), S. 99-102

[5] Choi, S. W. Some statistical properties and Zipf's law in Korean text corpus. *Journal of Quantitative linguistics* 7, 1, 2000.

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

[7] A. Dinu, L.P. Dinu. On the Syllabic Similarities of Romance Languages. In *A. Gelbukh (Ed.): CICLing 2005. LNCS 3406*, 785-788, 2005a.

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

[9] A. Dinu, L.P. Dinu. On the data base of Romanian syllables and some of its quantitative and cryptographic aspects. In Proceedings LREC 2006, Genoa, Italy, 1795-1798.

[10] *Dicţionarul ortografic, ortoepic şi morfologic al limbii române*. Ed. Academiei, Bucureşti, 1982.

[11] Elts, J., J. Mikk. Determination of optimal values of text. *Journal of quantitative linguistics* 3, 2, 1996.

[12] Fenk, A., G. Fenk-Oczlon. Menzerath's law and the constant flow of linguistic information. În *Contributions to quantitative linguistics*, eds. R. Köhler, B. B. Rieger. Kluwer Academic Publishers, Netherlands, 1993.

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

[14] Levelt, W.J.M., L. Wheeldon. Do speakers have access to a mental syllabary? *Cognition* 50, 239-269, 1994.

[15] 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.

[16] Marcus, S., Ed. Nicolau, S. Stati. *Introduzione alla linguistica matematica*, Bologna, Patron, 1971.

[17] Markov, A.A. An example of statistical investigation in the text of Eugen Onyegin illustrating coupling of tests in chain. În *Proceedings of the Academy of Science of St. Petersburgh* VI Series, 7, 153-162, 1913.

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

[19] 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

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

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