# ON THE QUANTITATIVE AND FORMAL ASPECTS OF THE ROMANIAN SYLLABLES

LIVIU P. DINU

**Abstract**. In this paper we investigate the syllable from two different points of view. In the first part we investigate the quantitative aspects of the syllable. Firstly, 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.

The second part is dedicated to formal approaches of the syllable. We propose and study a model of the graphical syllable, using Marcus contextual grammars. For this purpose we introduce two new variants of Marcus contextual grammars: total Marcus contextual grammar with total leftmost derivation and total Marcus contextual grammar with total leftmost derivation constrained by maximal use of selectors. A second formalization of the syllable is based on the presupposition that the syllabification is rather parallel than sequential one. The parallel manner of syllabification is realized by introducing some parallel extensions of insertion grammars. We use these grammars in an application to Romanian language syllabification.

## 1. INTRODUCTION

*Mathematical Linguistic*, as *the study of quantitative and formal aspects of language phenomenon* (Marcus *et al.*1971), has developed simultaneously in Europe and S.U.A in the late fifties.

Quantitative aspects of language were investigated long before the algebraic ones. Thus, there are records of all letters and diacritic symbols of Italian since the XIV$^{th}$–XVI$^{th}$ century; the Morse alphabet was inspired from the different statistic behavior of letters; in the XIX$^{th}$ century frequency dictionaries were edited and the beginning of the XX$^{th}$ century brings the first linguistically motivated studies which resulted in introducing the Markov models. The appearance of *Cours de linguistique generale* of Ferdinand de Saussure in 1916 resulted in placing the quantitative problems in the interest area of structuralist schools. One of the scientists who pleaded for the need of quantitative studies of the languages was the Czech scientist Vilem Mathesius (1929). The increasing numerous and complex

RRL, **LI**, *3–4*, p. 477–498, Bucureşti, 2006

**BDD-A246 © 2006 Editura Academiei**
**Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)**

studies in the field (Bloomfield – *Language*, 1933-, Trnka (1935), Troubetzkoy – *Principes de phonologie*, 1939–, Zipf, Yule, Ross, etc.) determined the organizers of the VI$^{th}$ international congress of linguistics, Paris, 1948 to create a *comity for investigating the quantitative aspects of linguistics*.

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) were the first who introduced the mathematical models in poetics. In 1978 Solomon Marcus realized a synthesis of the Romanian research in mathematical and computational linguistics till 1978. The papers presented are grouped in seven categories (statistical linguistics, algebraic linguistics, analytical models, generative models, mathematical and computational poetics, computational linguistics, applications of mathematical linguistics in science and art) and count over 500 titles and over 120 Romanian authors.

## 2. ON THE LINGUISTIC RESOURCES

In the last decade, the building of language resources and their 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. LRs also cover basic software tools for their acquisition, preparation, collection, management, customization and use and are used in many types of applications (from language services to e-learning and linguistic studies, etc.). 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 lots of data base of linguistic resources like tree banks, morphemes, dictionaries, annotated corpora, etc. In the last years, one of the linguistics structures that regained the attention of the scientific community from Natural Language Processing area was the syllable (Kaplan and Kay 1994, Levelt and Indefrey 2001, Muller 2002, Dinu 2003, Dinu and Dinu 2005a,b, 2006).
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, automatic 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 psycho-linguistic

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

investigation could greatly benefit from the existence of such a data base. These are some of the reasons which provided our motivation for creating a syllable data base for the Romanian language and to study its quantitative aspects. We must say that one of the first lexical resources regarding syllables was the database of Dutch syllables (Schiller *et al.* 1996). In the next section we will present in more detail the reasons for constructing a data base of syllables for Romanian language.

### 3. MOTIVATION. WHY THE SYLLABLE?

The first writing systems had the syllable as the basic unity, the first letter-based writing systems being used by the Greeks. In antiquity, in Greece and India, the syllable was discovered in poetics, when studying the metrics.

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).

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 and of the whole word. These aspects may have application in cryptography.

The psycholinguistic elements are situated inside the speech production area. Experiments revealed the presence of a library of articulator pre-compiled 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 articulator routines that is accessed during the process of speech production. Adjoining such syllabic gesture generates the spoken word and greatly reduce the computational cost of articulator programs.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

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.

## 4. MATHEMATICAL ASPECTS OF THE SYLLABLE

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

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 a 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 Table 1:

*Table 1*

Optimal length of words (Bamberge, Vanecek 1984, cf. Elts and Mikk 1996):

| The length of words | The reader's level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| in syllables | 1.62 | 1.68 | 1.72 | 1.8 | 1.88 | 1.91 | 1.99 | 2.08 | 2.11 |
| in letters | 6.16 | 6.39 | 6.39 | 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 (Table 2):

*Table 2*

| | Level 8 | Level 10 |
|---|---|---|
| Optimal length of words, measured in letters | 8.53 | 8.67 |
| Optimal length of sentences, measured in letters | 71.5 | 76 |

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

In order to properly study the quantitative aspects of the syllables from a given language or to investigate cognitive aspects of speech production one needs to build a syllable data base. One of the languages which has such a data base is Dutch. Analyzing this data base produced the following result (cf. Schiller *et al.* 1996): for Dutch, the first 500 *type syllables*, ordered after their frequency, (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*. These results support the *mental syllabary* thesis.

## 5. THE DATA BASE OF ROMANIAN SYLLABLES

As any selection, our synthesis from previous sections omitted some results, either from lack of space, or from lack of knowledge. In this section we present two studies regarding the Romanian syllables: first one was realized by Alexandra Roceric (1968) on a short corpus, and the second is our own study, realized on an much larger corpus (DOOM).

### 5.1 Roceric's phonostatistics

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

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 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 syllabical 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.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

### 5.2. A DOOM based study

We want to extend the research of Roceric to a larger corpus in order to obtain a comprehensive data base of Romanian syllables and to investigate theirs comportment.
Here are the two major problems we have 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 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*. Thus, the frequency factor is disregarded, each word from the dictionary being syllabified only once. This is not in accordance to the fact that words have different occurrence frequency in the spoken (or written) language, 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, journalism, 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 that 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 in two classes. The first one is formed by the rules which apply to a consonantal sequence of 1 to 5 (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 that apply to a sequence of vowels.
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. These rules do 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.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

## 6. THE ANALISE OF THE DATA BASE OF ROMANIAN SYLLABLES

The corpus we used (DOOM) contains *Nwords* = 74276 words. We automatically syllabified the words using an algorithm 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 possibilities of combination of the syllable (i.e. which are the syllables which can follow it and can be followed by it).

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

1. We identified *NStype* = 6496 (*type syllables*) in Romanian language. The total number of syllables (*token syllables*) is *NStoken* = 273261. So, the average length of a word measured in syllables is *Lwordssyl* = *Nstoken/Nwords* =273261/74276 = 3,678.
2. The 74276 words are formed of *Nletters* = 632702 letters. So, the average length of a word measured in letters is *Lwordslet* = *Nletters=/Nwords* =632702/74276 = 8,518.
3. In order to characterize the average length of a syllable measured in letters we investigate two cases:
    a. the average length of the *token syllables* measured in letters is: *Lsyltoken* = *Nletters/NStoken* = 632706/273261 = 2,315;
    b. The *type syllables* are formed of *NTletters* = 24406 letters. Thus, the average length of a *type syllable* measured in letters is *Lsyltype =Ntletters/NStype* = 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*:

*Table 3*

CV structure for the type syllables:

| 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% |

*(to continue)*

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

Table 3 *(continued)*

| | | |
|---|---|---|
| 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% |

b. for the *token-syllables*:

*Table 4*

CV structure for the token syllables:

| 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% |

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 *NStype*) have 137662 occurrences, i.e. 50.03% of *NStoken*.
6. The most frequent 200 syllables cover 76% of *NStoken*, the most frequent 400 cover 85% of *NStoken* and the most frequent 500 syllables (i.e. 7.7% of *NStype*) cover 87% of *NStoken*. Over this number, the percentage of covering rises slowly.
7. The first 1200 syllables in there frequency order cover 95% of *NStoken*.
8. 2651 syllables of *NStype* occur only once (hapax legomena).
9. 5060 syllables (i.e. 78%) of *NStype* occur less then 10 times. These syllables represent 11960 syllables (4% of *NStoken*).

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

10. 158941 syllables (58% of *NStoken*) are formed of 2 letters; the syllables formed of 3 letters represent 27% of *NStoken*, those formed of 1 letter represent 9% of *NStoken* and those formed of 4 letters represent 6% of *NStoken*.
11. We computed the entropy of syllables, using the formula:

$$Hsyl = -\Sigma_{i=1,\,6496}\, p_i \log_2 p_i \quad (1)$$

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 *pi* is computed as the ratio between the total frequency of the syllable situated on the *i*-th position and the total number of occurrences *NStoken*. Thus, we obtained that the value of the syllable entropy is:

$$Hsyl = 8{,}621 \quad (2)$$

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

$$Hsyl = -\Sigma_{i=1,\,56}\, p_i \log_2 p_i \quad (3)$$

where *pi* 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 the value:

$$Hsyl = 2{,}30 \quad (4)$$

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

## 7. THE LAWS OF CHEBANOW, MENZERATH AND FENK FOR ROMANIAN SYLLABLES

In this section we investigate the behaviour of Romanian syllables related to these three laws.

### 7.1. Chebanow's law

One of the most 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 the words. Denoting by $F(n)$ the frequency of a word having $n$ syllables and by

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

$$i = \frac{\Sigma nF(n)}{\Sigma F(n)} \tag{5}$$

the average length (measured in syllables) of the words, Chebanow proposed the following law between the average $i$ and the probability of occurrences $P(n)$ of the words having $n$ syllables:

$$P(n) = \frac{(i-1)^{n-1}}{(n-1)!} e^{1-i} \tag{6}$$

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 the length (in syllables) of the words (Fig. 2):

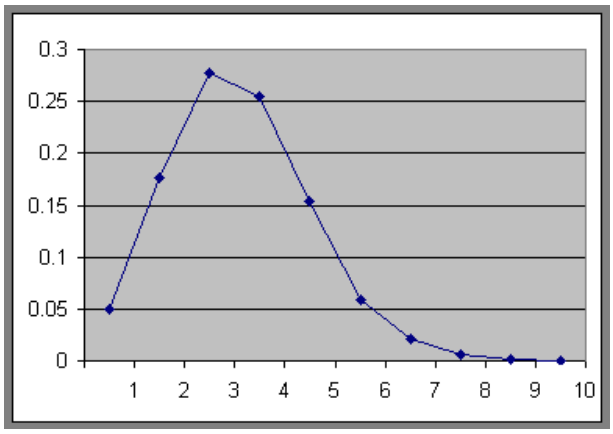



Fig. 1 – The probability distribution of the length of the words

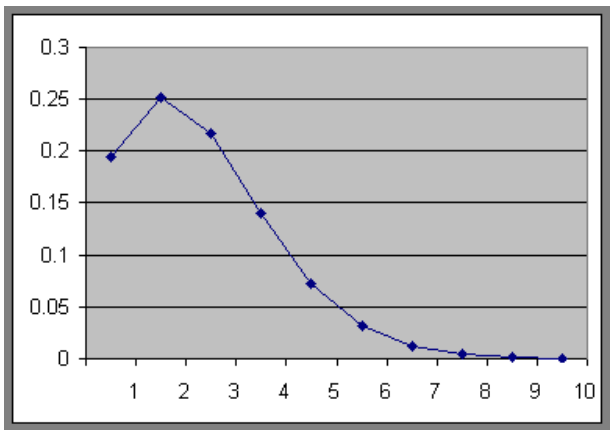



Fig. 2– The Poisson distribution of the length (in syllables) of the words (parameter equal to 2,678)

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

**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 Fig. 1 we represented the following Poisson's distribution (the average length of words is 3,678, so we have to use the value 3.678-1=2.678, cf. Chebanow's law) :*

$$P(n) = \frac{2,678^n}{n!} e^{-2,678} \tag{7}$$

### 7.2. 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. Fig. 3 shows that the law is satisfied.

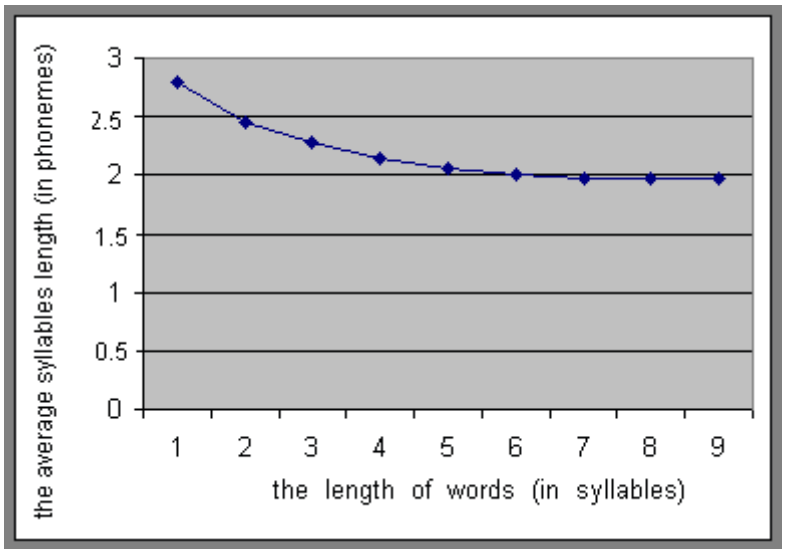



Fig. 3– Menzerath's law

### 7.3. Fenk's law

Fenk (1993) observed also that the bigger the length of a word, measured in phonemes, the lesser the length of its constituent syllables, measured in phonemes. We checked this correlation and the Fig. 4 confirms the first Fenk's law:

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

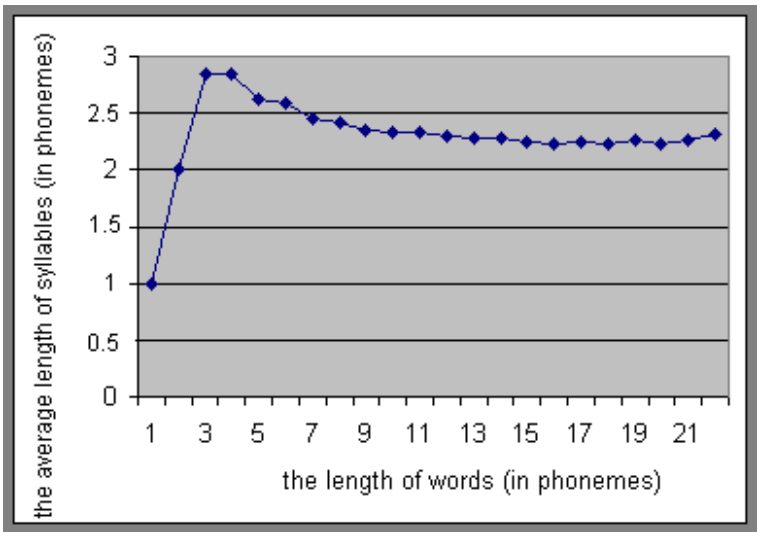



Fig. 4 – The first Fenk's law

**8. FORMAL APPROACHES OF THE SYLLABLES**

It is well-known that Noam Chomsky introduced his formal grammars as tools for formalizing the syntax of natural languages, as he explicitly stated (Chomsky 1957): *"the main problem of immediate relevance to the theory of language is that of determining where in the hierarchy of devices the grammars of natural languages lie. It would, for example, be extremely interesting to know whether it is in principle possible to construct a phrase structure grammar for English."*

On the other hand, 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 formal 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 the last three decades, the formal devices where used to analyze not only the syntax, but also the morphology, the phonology and various other linguistic fields and some mathematical models of the syllable were proposed.

Based mostly on set theory, the universal phonological model of the syllable is introduced by Theo Vennemann (1978). Koskenniemi (1983) proposed a computational model to recognize and product the morphological and phonological word-form (two-level morphology). Bird and Ellison (1994) used finite automata to model the rules of phonological segmentation. Kaplan and Kay (1994) show how the algebra of regular relations, with their corresponding automata, can be

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

used to *"establish a solid basis for computation in the domain of phonological and orthographic systems"*. Bird and Klein (1994) used the formal resources of HPSG to treat in a rigorous fashion various phonological constructs. Karin Muller (2002) developed a probabilistic syllable model (based on context free grammars) for German and English. Her model can be used for syllabification and grapheme-to-phoneme conversion in a speech system. Based on the similarity between the syllabification of a word and the generation of a word by a total Marcus contextual grammar, in (Dinu 2003) we proposed a contextual model of syllabification, using some extensions of contextual grammars. In (Dinu *et al.* 2004) we introduced the syllabic grammars and show how the syllabification can be modeled by *go-through automata*.

## 9. CONTEXTUAL APPROACHES TO THE SYLLABLE

One of the main problems of structural linguistics is the segmentation, i.e. the modality to divide a linguistic construct into its constituents, on different levels (e.g. phonemes, morphemes, etc.). Lately, many people analyzed the modality of segmentation in syllables of words, with direct applications in the speech synthesis and recognition.

In formal language theory, most of the generative mechanisms investigated are based on the *rewriting* operation. Several other classes of mechanisms, whose main ingredient is the *adjoining* operation, were introduced along the time. The most important of them are *the contextual grammars* (Marcus 1969), *the tree adjoining grammars* (TAG) (Joshi *et al.* 1975) and *the insertion grammars* (Galiukschov 1981), all three of them introduced with linguistic motivations.

In the next sections we use the contextual grammars and the insertion grammars to propose a sequential and a parallel manner of syllabification of words, respectively.

### 9.1. Contextual grammars

Contextual grammars have their roots in the development of structural linguistics and in the need of avoiding some of the shortcomings of the already existing generative devices.

Contextual grammars were introduced by Marcus (1969), as "intrinsic grammars", without auxiliary symbols, based only on the fundamental linguistic operation of inserting words in given phrases, according to certain contextual dependencies. In (Marcus 1997) S. Marcus has explained the circumstances and the motivation of introducing contextual grammars: *"... generative grammars are a rupture from the linguistic tradition of the first half of XX$^{th}$ century, while analytical models are just the development, the continuation of this tradition. It was natural to expect an effort to bridge this gap. This effort came from both parts*

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

*and, as we shall see,contextual grammars are a component of this process."*; he continues: *"Contextual grammars have their origin in the attempt to transform some procedures developed within the framework of analytical models into generative devices"* and *"The concept of contextual grammars takes into account the capacity of any string to select a class of preferential contexts. However, this capacity is only a part of a more comprehensive phenomenon, the duality between strings and contexts."*

More precisely, a contextual grammar produces a language starting from a finite set of words and iteratively adding contexts to the currently generated words, according to a selection procedure: each context associates with it a selector, a set of words; the context is adjoined to any occurrence of such a selector in the string to be derived.

Up to now, the contextual grammars were investigated mainly from a mathematical point of view and a series of important results are obtained; see (Marcus, Martin-Vide and Păun 1998; Martin-Vide, Mateescu, Miguel-Verges and Păun 1995; Păun, 1997) and their references. Recently some efficient parsers have been constructed (Gramatovici 1998; Harbusch 2000). In a series of papers (Marcus, Martin-Vide and Păun 1998; Martin-Vide 1997), some types of contextual grammars were used as generative models of natural languages, at their syntactic level. We used the contextual grammars in the investigation of syllabic segmentation.

### 9.2. A contextual approach to the syllable

In this section we shortly present the contextual approaches of the syllable and of the syllabification (Dinu 2003).

Suppose now that a phrase is generated by a contextual grammar. This means that each step of the derivation also corresponds to a correct phrase. Similarly, during the syllabification of a word, we can assume that a correct cutting was obtained whenever we stopped. This similarity made contextual grammars an attractive model for syllabification.

**Definition 1.** (Păun, 1997) *A total Marcus contextual grammar is a system* $G = (V, A, C, \varphi)$ *,where V is an alphabet, A is a finite language over V (the axioms), C is a finite subset of V x V (the contexts) and* $\varphi : V \times V \times V \longrightarrow P(C)$ *(the choice function)*

The language generated by $G$ is:

$$L(G) = \{x \in V^* \mid w \xrightarrow{*} x, \text{ for } w \in A\} ,$$

where "$\xrightarrow{*}$" is the reflexive and transitive closure of "$\rightarrow$", given by:
$x \rightarrow y$ iff $x = x1x2x3$, $y = x1ux2vx3$ for $x1, x2, x3 \in V^*$, and $<u, v> \in C$ such that $<u, v> \in \varphi(x1, x2, x3)$.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

Consider the Romanian alphabet *RO={*a, ă, â, , b, c, d, e, f, g, h, i, î, j, k, l, m, n, o, p, q, r, s, ş, t, ţ, u, v, w, x, y, z} and consider a nontrivial partition *RO= Vo* ∪ *Co*, where *Vo={a,ă, â,, e, i, î, o, u, y}* and *Co={b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, ş, t, ţ, v, w, x, z}*, i.e., *Vo* and *Co* are the Romanian vowels and the Romanian consonants, respectively.

We will say that *a word over RO is regular if it contains no consecutive vowels.*

In order to generate all the Romanian syllables which appear in regular words, and only them, we propose the grammar *Gsyl* = (*Vsyl,Asyl, ;Csyl,* $\varphi_{syl}$), whose components are:

1. *Vsyl* = *RO* ∪ {$}, where "$" is a new symbol that is not in *RO*; "$" is the *syllable boundary marker*.
2. *Asyl* is the set of the regular words over *RO*. *Asyl* is finite since the set of all words in a natural language is finite.
3. $Csyl = \{<\lambda,\lambda>,<\lambda,\$>,<\$,\lambda>\}$
4. $\varphi_{syl}$ is defined based on the syllabification rules of the Romanian languages (DOOM, 1982).
   a. $\varphi_{syl}(\alpha v_1, c, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c \in Co,\ v_{1,2} \in Vo$ (i.e. in the case of a consonant between two vowels, the syllabification is done before the consonant)
   b. $\varphi_{syl}(\alpha v_1, c_1c_2, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c_1c_2 \in \{ch, gh\},\ or$ $(c_1,c_2) \in \{$b, c, d, f, g, h, p,t$\} \times \{l, r\}$
   c. $\varphi_{syl}(\alpha v_1c_1, c_2, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c_1c_2 \notin \{ch, gh\},\ and$ $(c_1,c_2) \notin \{$b, c, d, f, g, h, p,t$\} \times \{l, r\}$
   d. $\varphi_{syl}(\alpha v_1c_1, c_2c_3, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c_1c_2c_3 \notin \{$lpt, mpt, mp ţ, nc ş, nct, nc ţ, ndv, rct, rtf, stm$\}$
   e. $\varphi_{syl}(\alpha v_1c_1, c_2, c_3v_2\beta) = \{<\lambda,\$>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c_1c_2c_3 \in \{$lpt, mpt, mpt,, nc,s, nct, nct,, ndv, rct, rtf, stm$\}$
   f. $\varphi_{syl}(\alpha v_1c_1, c_2c_3c_4, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,\ c_2c_3c_4 \notin \{$gst, nbl$\}$
   g. $\varphi_{syl}(\alpha v_1c_1c_2, c_3c_4, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,$ $c_2c_3c_4 \in \{$gst, nbl$\}$
   h. $\varphi_{syl}(\alpha v_1c_1c_2, c_3c_4c_5, v_2\beta) = \{<\$,\lambda>\} if \ \ \alpha,\beta \in V_{syl}^*,$ $c_1c_2c_3c_4c_5 \in \{$ptspr,stscr$\}$
   i. $\varphi_{syl}(x_1, x_2, x_3) = \{<\lambda,\lambda>\}, otherwise$

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

The language generated by *Gsyl* is:

$$L(G_{syl}) = \{x \in V_{syl}^* \mid w \xrightarrow{*} x \ \ for \ \ w \in A_{syl}\}$$

and it contains all possible ways of syllabification regular words (for example, the language contains the word *lingvistica* and all its possible syllabifications: *lin$gvistica, lingvis$tica, lingvisti$ca, lingvis$ti$ca, lin$gvisti$ca, lin$gvis$tica, lin$gvis$ti$ca*).

We introduce the set *Syl* as follows:

$$Syl = \{x \in (V_{syl} \setminus \$)^+ \mid \exists \alpha, \beta \in (V_{syl})^* \ \ such \ \ that \ \ \alpha x \beta \in L(G_{syl})$$

$$and \ \ x \Rightarrow y \ \ implies \ \ x = y\}$$

This definition allows us to define the syllable as it follows:

**Definition 2.** *A segment* $syl \in \{Co \cup Vo\}^*$ *is a syllable iff* $syl \in Syl$.

**Remark 3**. *In most of the natural languages there are words which have different syllabifications. For Romanian words, the only words which can have two different syllabifications are the words ending in "i" (e.g. ochi (noun) and o$chi (verb)) (Petrovici, 1934). The syllabification of such a word depends on whether the final "i" is stressed or not. If the final "i" is stressed, the rules a)-i) are applied , else the final "i" is considered as a consonant and then the same rules are applied.*

**Remark 4**. *Inside a graphical non regular word, in a sequence of 2, 3, 4 or 5 vowels it is difficult to distinguish between a vowel and a semivowel. In order to cut into syllables such a word we have tried to extract a set of rules based on the context in which the sequence appears. Thus, we notice that the same group of vowels has an identical behavior(regarding the syllabification of words which contains it) depending on certain letters which precede and/or succeed it (Dinu, 1997). Once we have founded a set of rules which characterize the behavior of a sequence of vowels, we use it to extend the grammar Gsyl. We have obtained a set of rules which characterize the behavior of some sequences of vowels, the rest of them being under construction.*

**Remark 5.** For a word $w \in V_{syl}^*$ there may be two different decompositions of *w*, $w = x1x2x3$ and $w = y1y2y3$, such that using direct derivation we can obtain two different words, $w = x1x2x3 \Rightarrow x1ux2vx3 = w1$ and $w = y1y2y3 \Rightarrow y1uy2vy3 = w_2$, with $w1 \neq w_2$. In other words, the syllabification may be done anywhere inside the word, the only condition being that the cutting should be correct.

**Example 1**. Consider the word *lingvistica*. We may have the follow direct derivations:

1. *lingvistica* $\Rightarrow$ *lin*$*gvistica*
2. *lingvistica* $\Rightarrow$ *lingvisti*$*ca*

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

To avoid these situations, we shall impose that the cutting to be always done at the leftmost position. For this purpose we have considered a series of constraints of the derivation relation defined with respect to a total contextual grammar, called *total leftmost derivation*. By using it, contexts are introduced in the leftmost possible place.

## 10. A PARALLEL APPROACH TO THE SYLLABLE

The previous model, like most of the formal models of syllabication, are treated in a sequential manner. It is highly conceivable that our brain works in a distributed parallel manner when producing phrases in a natural language. Our belief is that, based on a set of rules (innate or acquired from experience) the brain uses a parallel mechanism to syllabicate the words. This is in prosecution of some cognitive theory of speech production (Levelt and Indefrey 2001) and could enable the brain to reduce the duration of speech production. In (Dinu and Dinu 2005a) we proposed a parallel manner of syllabification, introducing some parallel extensions of insertion grammars.

### 10.1. Insertion grammars

The basic operation in contextual grammars is the adjoining of contexts, depending on the string bracketed by the two added strings; in Chomsky context-sensitive grammars, a symbol is rewritten by a string, depending on a context. The insertion grammars were introduced by Galiukschov in 1981 and are an intermediate model: strings are inserted in a context. Again the basic operation is the adjoining of strings, as in contextual grammars, not rewriting, as in Chomsky grammars, but the operation is controlled by a context, as in context-sensitive grammars.

**Definition 3.** (Păun, 1997) *An insertion grammar is a triple* $G = (V, A, P)$*, where V is an alphabet, A is a finite language over V, and P is a finite set of triples of strings over V. The elements in A are called axioms and those in P are called insertion rules.*

*The meaning of a triple* $(u,x,v) \in P$ *is: x can be inserted in the context* $(u, v)$*. Specifically, for* $w, z \in V^*$*, we write* $w \Rightarrow z$ *iff* $w = w1uvw2$*;* $z = w1uxvw2$*, for* $(u, x, v) \in P$ *and* $w1, w2 \in V^*$.

The language generated by G is defined by:

$$L(G) = \{z \in V^* \mid w \xrightarrow{*} z \ \ for \ \ w \in A\}.$$

In order to propose a parallel syllabification, we introduced two parallel extensions of insertion grammars.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

**Definition 4.** *Let $G = (V, A, P)$ be an insertion grammar. We define the parallel derivation denoted $\Rightarrow_p$, by:*

$w \Rightarrow_p z$ *iff* $w = w_1 w_2 ... w_r$, *for some* $r \geq 2$, $z = w_1 x_1 w_2 x_2 ... x_{r-1} w_r$,

*and, for all* $1 \leq i \leq r-1$,

*there is* $(u_i, x_i, v_i) \in P$ *and* $\alpha_i, \beta_i \in V^*$ *such that* $w_i x_i w_{i+1} = \alpha_i u_i x_i v_i \beta_i$,

*and* $w_i = \alpha_i u_i$, $w_{i+1} = v_i \beta_{i+1}$.

**Remark 6.** *For usual derivation $\Rightarrow$ we use one selector-pair, with no restriction; in parallel derivations the whole string is decomposed into selectors.*

**Definition 5.** *For an insertion grammar $G = (V; A; P)$ the parallel derivation with maximum use of insertions (in short, we say maximum parallel derivation), denoted $\Rightarrow_{pM}$, is the parallel derivation applied with maximum possible insertions.*

**Remark 7.** *The main difference between parallel derivation and maximum parallel derivation with respect to an insertion grammar is that in the former we can insert any number of strings in a derivation step and in the later we insert the maximum possible number of strings in a derivation step.*

The family of languages generated by an insertion grammar in the mode $\alpha \in \{p, pM\}$ is denoted by $INS_p$, $INS_{pM}$, respectively.

## 10.2 On the syllabification of Romanian words via parallel insertion grammars

In this section we use the insertion grammars and the maximum parallel insertion derivation to propose a parallel manner of syllabification of words.

Consider the Romanian alphabet *RO={*a, ă, â, b, c, d, e, f, g, h, i, î, j, k, l, m, n, o, p, q, r, s, ş, t, ţ, u, v, w, x, y, z} and its partition in vowels and consonants: *RO= Vo* ∪ *Co*, where *Vo={a,ă, â,, e, i, î, o, u, y}* and *Co={b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, ş, t, ţ, v, w, x, z}*.

With respect to the above definitions, an insertion grammar for syllabification of Romanian regular words is *Gsyl* = (*Vsyl,Asyl, Psyl*), whose components are:

1. *Vsyl* = *RO* ∪ {$}, where "$" is a new symbol that is not in *RO*; "$" is the *syllable boundary marker*.
2. *Asyl* is the set of the regular words over *RO*.
3. *Psyl* = *C*1 ∪ *C*2 ∪ *C*3 ∪ *C*4 ∪ *C*5 ∪ *C*6 ∪ *C*7 ∪ *C*8, where:
   a. $C_1 = \{(v_1, \$, cv_2) \mid c \in Co,\ v_{1,2} \in Vo\}$
   b. $C_2 = \{(v_1, \$, c_1c_2v_2) \mid v_{1,2} \in Vo, c_1c_2 \in \{ch, gh\},$
      *or* $(c_1, c_2) \in \{b, c, d, f, g, h, p, t\} \times \{l, r\}\}$

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

c. $C_3 = \{(v_1c_1, \$, c_2v_2) \mid v_{1,2} \in Vo, c_1c_2 \notin \{ch, gh\},$
*and* $(c_1, c_2) \notin \{b, c, d, f, g, h, p, t\} \times \{l, r\}\}$

d. $C_4 = \{(v_1c_1, \$, c_2c_3v_2) \mid v_{1,2} \in Vo, c_1c_2c_3 \notin \{$lpt, mpt, mpț, ncș, nct, ncț, ndv, rct, rtf, stm$\}\}$

e. $C_5 = \{(v_1c_1c_2, \$, c_3v_2) \mid v_{1,2} \in Vo, c_1c_2c_3 \in \{$lpt, mpt, mpț, ncș, nct, ncț, ndv, rct, rtf, stm$\}\}$

f. $C_6 = \{(v_1c_1, \$, c_2c_3c_4v_2) \mid v_{1,2} \in Vo, c_2c_3c_4 \notin \{gst, nbl\}\}$

g. $C_7 = \{(v_1c_1c_2, \$, c_3c_4v_2) \mid v_{1,2} \in Vo, c_2c_3c_4 \in \{gst, nbl\}\}$

h. $C_8 = \{(v_1c_1c_2, \$, c_3c_4c_5v_2) \mid v_{1,2} \in Vo, c_1c_2c_3c_4c_5 \in \{ptspr, stscr\}\}$

**Example 2.** *Consider the word lingvistica. We may have the following parallel derivations:*

- *A parallel derivation: lingvistica* $\Rightarrow$ *lin$gvisti$ca, where:*
  - *a.* *i=1:* $w_1x_1w_2 = \alpha_1u_1x_1v_1\beta_1$, *with* $(u_1, x_1, v_1) \in C_4$: $\alpha_1 = l,\ u_1 = in,\ x_1 = \$,\ v_1 = gvi,\ \beta_1 = sti$
  - *b.* *i=2:* $w_2x_2w_3 = \alpha_2u_2x_2v_2\beta_2$, *with* $(u_2, x_2, v_2) \in C_3$: $\alpha_2 = gvist,\ u_2 = i,\ x_2 = \$,\ v_2 = ca,\ \beta_2 = \lambda$
- *Maximal Parallel derivation: lingvistica* $\Rightarrow$ *lin$gvis$ti$ca*
  - *a.* *i=1:* $w_1x_1w_2 = \alpha_1u_1x_1v_1\beta_1$, *with* $(u_1, x_1, v_1) \in C_4$: $\alpha_1 = l,\ u_1 = in,\ x_1 = \$,\ v_1 = gvi,\ \beta_1 = s$
  - *b.* *i=2:* $w_2x_2w_3 = \alpha_2u_2x_2v_2\beta_2$, *with* $(u_2, x_2, v_2) \in C_3$: $\alpha_2 = gv,\ u_2 = is,\ x_2 = \$,\ v_2 = ti,\ \beta_2 = \lambda$
  - *c.* *i=3:* $w_3x_3w_4 = \alpha_3u_3x_3v_3\beta_3$, *with* $(u_3, x_3, v_3) \in C_1$: $\alpha_3 = t,\ u_3 = i,\ x_3 = \$,\ v_3 = ca,\ \beta_3 = \lambda$

## 11. COGNITIVE ASPECTS OF THE CONTEXTUAL MODELS OF THE SYLLABLE

In a cognitive perspective, the simple operation of adjoining might be closer than rewriting to the way our brain may work when constructing a phrase. It is hard to imagine our brain using auxiliary intermediate phrase of a non-terminal type. Instead, it looks more natural to start with a collection of well-formed phrases, maybe acquired from experience, and to produce new well-formed ones by adding further words, in pairs that can observe dependencies and agreements, and in accordance with specified selectors, which can ensure the preservation of

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

grammaticality (Martin-Vide 1997; Marcus, Martin-Vide and Păun 1998). It seems that this hypothesis is in concordance with one of the major theories developed for speech production (Levelt and Indefrey 2001).

In normal speech we produce words at rates 2 to 4 per second. The theory proposed consists of two major processing component. The first component deals with *lexical selection*. It is the mechanism that, given semantic input (some state of affairs to be expressed), selects one appropriate lexical item from the mental lexicon. The second component deals with *form encoding*. It computes the articulator gestures needed for the articulation of the selected items.

The first step here is the retrieval of the target item's phonological code, an abstract string of phonological segments. The next operation is syllabification. Segments are incrementally concatenated (adjoining) to form syllables. Segmental concatenation in syllabification runs at a rate of about 25 milliseconds per segment. The final step in form encoding is *phonetic encoding*, the retrieval of articulator scores for each of the incrementally generated syllables. The theory of Levelt and Wheeldon (1994) assumes the existence of a *mental syllabary*: for frequently used syllables there is a library of articulator routines that is accessed during the process of speech production. The adjoining of such syllabic gesture combined with a parallel manner of syllabification greatly reduces the computational cost of generating the spoken words.

## 12. CONCLUSIONS AND FUTURE WORKS

In the first part of this paper we have presented some quantitative observations obtained from the analise of the first data base of Romanian syllables. We also computed the entropy of the syllables and the entropy of the syllables w.r.t. the consonant-vowel structure 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 1996, Choi 2000).

In the second part of the paper we have investigated the contextual grammars as generative models for the natural language. We introduced some constraints to the derivation relation, obtaining new contextual grammars. Using the languages generated by these grammars we proposed a contextual model of the syllable.

From the cognitive point of view, a model based on contextual grammar seems close to the way the brain operates when it produces speech.

The development of our contextual model for syllabification was based on the Romanian rules of syllabification, but it can be adjusted for any language.

In some future work we hope to be able to present results obtained by analyzing a corpus of spoken Romanian language other than the one we used (DOOM) and compare them to the results in this paper.

BDD-A246 © 2006 Editura Academiei
Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)

**Acknowledgements:** MEdC-ANCS, CNCSIS and CNR-NATO have supported this research.

REFERENCES

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

Altmann, G., 1980, "Prolegomena to Menzerath's law", in R. Grotjahn (ed.), *Glottometrika 2*, 1–10, Bochum.

Altmann, G., 1993, "Science and linguistics", in R. Kohler, B. B. Rieger (eds.), *Contributions to quantitative linguistics*, Kluwer Academic Publishers, Netherlands.

Bird, S., T. M. Ellison, 1994, "One-level phonology: Autosegmental representations and rules as finite automata", *Computational Linguistics,* 20, 55–90.

Chebanow, S.G., 1947, "On conformity of language structures within the Indoeuropean family to poisson's law", *Comptes rendus de l'Academie de science de l'URSS*. 55, S. 99–102

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

Dinu, L.P., 1997, "The alphabet of syllables with applications in the study of rime frequency", *Analele Universităţii din Bucureşti,* XLVI, 39–44.

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

Dinu, L.P., 2004, *Metode formale şi de clasificare în lingvistica matematică şi computaţională*, Bucureşti, Editura Universitatii din Bucureşti.

Dinu, L. P., A. Dinu, 2005a, "On the Syllabic Similarities of Romance Languages", in: *A.* Gelbukh (ed.), *CICLing 2005. LNCS 3406*, 785–788.

Dinu, L. P., A. Dinu, 2005b, "A parallel approach to syllabification", in: A. Gelbukh (ed.), *CICLing 2005. LNCS 3406*, 83–87.

Dinu, L. P., A. Dinu, 2006, "On the database of the Romanian syllables and some of its quantitative and cryptographic aspects", in *Proceedings LREC 2006, Genoa, Italy*, 1795–1799.

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

*Dicţionarul ortografic, ortoepic şi morfologic al limbii române*, 1982, Bucureşti, Editura Academiei.

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

Fenk, A., G. Fenk-Oczlon, 1993, "Menzerath's law and the constant flow of linguistic information", in: R. Kohler, B. B. Rieger (eds) *Contributions to quantitative linguistics*, Netherlands, Kluwer Academic Publishers.

Galiukschov, B. S., 1981, "Semicontextual grammars", (in Russian), *Mat. logica i mat. ling.*, Kalinin Univ. 38–50.

Gramatovici, R., 1998, "An efficient parser for a class of contextual languages", *Fundamenta Informaticae*, 33, 211–238.

Harbusch, K., 2000, "Parsing contextual grammars with linear, regular and context free selectors", in: C. Martin-Vide, V. Mitrana (eds), *Words, Sequences, Grammars, Languages, Where Biology, Computer Science and Mathematics meet II*, Springer, London, UK.

Herdan, G., 1964, *Quantitative Linguistics*, Butterworths.

Joshi, A.K., L.S. Levy, M. Takahashi, 1975, "Tree adjoining grammars", *J. Computer System Sci.*, 19, 136–163, 1975.

Kaplan, R. M., M. Kay, 1994, "Regular models of phonological rule systems", *Computational Linguistics*, 20, 3, 331–379.

Levelt, W. J. M., L. Wheeldon, 1994, "Do speakers have access to a mental syllabary?", *Cognition,* 50, 239–269.

**BDD-A246 © 2006 Editura Academiei**
**Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)**

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

Marcus, S., 1969, "Contextual grammars", *Revue Roumaine de Mathémathiques Pures Appliqueés,* 14, 69–74.

Marcus, S., 1978, "Mathematical and computational linguistics and poetics", *Revue Roumaine de Linguistique*, XXIII, 559–588.

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

Marcus, S., 1997, "Contextual grammars and natural languages", in G. Rozenberg, A. Salomaa (eds) *Handbook of Formal Languages*, vol. 2, Springer-Verlag Berlin Heidelberg.

Marcus, S., C. Martin-Vide, Gh. Păun, 1998., "Contextual Grammars as Generative Models of Natural Languages", *Computational Linguistics,* 24, 2, 245–265.

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

Martin-Vide, C., 1997, "Natural Computation for Natural Languages", *Fundamenta Informaticae,* 31, 117–124.

Martin-Vide, C., A. Mateescu, J. Miguel-Verges, Gh. Păun, 1995, "Contextual Grammars with maximal, minimal and scattered use of contexts", in: M. Koppel, E. Shamir (eds) *Proc. of the Fourth Bar-Ilan Symp. on Foundations of AI, BISFAI '95* , Jerusalem, 132–142.

Mateus, M. H., E. D'Andrade, 1998, "The syllable structure in European Portuguese", *D.E.L.T.A.* 14, 1, 13–32.

Menzerath, P., 1954, "Die Architektonik des deutschenWortschatzes", in *Phonetische Studien*, Heft 3. Bon, Ferd. Dummlers Verlag.

Muller, K., 2002, *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

Nicolau, E., 1962, "Langage et strategie", *Cahiers de linguistique théorique et appliquée*, 1, 153–179.

Păun, Gh., 1997, *Marcus Contextual Grammars*, Kluwer.

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

Roceric-Alexandrescu, A., 1968, *Fonostatistica limbii române*, Bucureşti, Editura Academiei.

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

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

Saussure, F., 1998, *Curs de lingvistică generală*, (traducere I. I. Tarabac), Iaşi, Editura Polirom.

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

Vennemann, T., 1978, "Universal syllabic phonology", *Theoretical Linguistics,* 5, 2–3, 175–215.

**BDD-A246 © 2006 Editura Academiei**

**Provided by Diacronia.ro for IP 216.73.216.143 (2026-11-18 22:06:38 UTC)**