ItGraSyll: A Computational Analysis of Graphical Syllabification and Stress Assignment in Italian

Liviu P. Dinu[1,3,*], Bogdan Iordache[1,3], Bianca Guita[3], Simona Georgescu[2,3] and Alina Cristea[3]

[1]University of Bucharest, Faculty of Mathematics and Computer Science, Romania
[2]University of Bucharest, Faculty of Foreign Languages and Literatures, Romania
[3]Human Language Technologies Research Center, Bucharest, Romania

Abstract

In this paper we build a dataset of Italian graphical syllables (called ItGraSyll). We perform quantitative and qualitative analyses on the syllabification and stress assignment in Italian. We propose a machine learning model, based on deep-learning techniques, for automatically inferring syllabification and stress assignment. For stress prediction we report 94.45% word-level accuracy, and for syllabification we report 98.41% word-level accuracy and 99.82% hyphen-level accuracy.

Keywords

syllabification, stress assignment, Italian,

1. Introduction

Word syllabification and syllable analysis are two related issues of great importance in the study of language (written or spoken). These topics have attracted a large category of researchers, from pure linguists, in phonetics, to psycholinguists, computer scientists, speech therapists, etc. Thus, the syllable plays an important role in language learning and acquisition, speech recognition, speech production [1, 2], language similarity [3], in text comprehensibility (Kincaid-Flesch formula [4]), in speech therapy, in poetry analysis [5, 6], etc. Each language has its own way of grouping sounds into syllables and its own rules for dividing words into syllables. Linguistically, the syllable represents "the smallest phonetic trance likely to receive an accent and only one" [7], and the syllabic cut is seen by De Saussure [8] on the border between the implosion and the explosion of the spoken sound: "If in a chain of sounds one goes from implosion to explosion, one obtains a particular effect which is the indication of the boundary of the syllable".

The analysis of the words' syllabic structure also plays an important part in historical linguistics [9], not only in diachronic phonetics and phonology, but also in lexicology. Romance comparative linguistics, in particular, still needs a detailed overview of this aspect, as syllable, segmentation and prosody can give strong account on phonetic changes that haven't been explained yet. The "prosodic revolution" [10] from Latin to the Romance languages – including syncope (the loss of an intermediate syllable) and apocope (the loss of the final syllable) at a large scale – has led to major changes, but their weight is different from one idiom to another: while the Western Romance languages manifest highly evident differences from the Latin phonological and prosodic system, and the Eastern languages are considered to be most conservative from this point of view, Italian seems to be in between [10]. On the other hand, in Latin, the relation between stress and quantity grew stronger, thus short stressed vowels progressively gained length. It is noteworthy that this situation is best preserved in Italian, and not in the Eastern Romance idioms: thus, in Italian stress cannot skip a heavy penultimate syllable, and stress cannot fall further back than the antepenultimate syllable, a twofold characteristic feature of the Latin prosodic system. This is why we are taking Italian as a starting point for a larger-scale study, oriented towards all Romance languages. The main difference between Latin and its modern descendants is that Latin stress was quantity- sensitive, leading thus to the following rule: in polysyllabic words, stress fell on a heavy penultimate (meaning, containing a long vowel), otherwise on the antepenultimate. Due to the collapse of vowel quantity as a distinctive feature in the vocalic system, no Romance language has retained the Latin stress rule as such [10]. As, from a statistic point of view, the greatest part of the Romance lexicon is represented by penultimate stressed words, a basic automatic mechanism would assign penultimate stress by default, whereas for both final and antepenultimate stress, the machine (as well as, not in a few cases, non-native speakers) would need further specification. As a consequence of the loss of Latin vowel quantity, Romance stress has ceased to be completely predictable. That is, partially, why in the majority of the traditional Romance compara-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
ldinu@fmi.unibuc.ro (L. P. Dinu);
iordache.bogdan1998@gmail.com (B. Iordache);
bianca.guita@s.unibuc.ro (B. Guita);
simona.georgescu@lls.unibuc.ro (S. Georgescu);
alinaciobanu20@gmail.com (A. Cristea)

tive or historical grammars, there is no specific section devoted to syllabification [11], or, if there is, it focuses either on general prosodic features [12], or on the vowel evolution depending on its presence in an open or closed syllable [13]. The lack of a section dedicated to syllabification is also common in the historical grammars of Italian [14, 11, 15]. We will focus in this research only on written form of words, so we will investigate only the graphical syllabification and stress. By focusing on the graphical syllabification and stress in Italian, we aim to take a step forward towards the complete evaluation of the prosodic changes that took place in the transition from Latin to the Romance languages, and their influence on the Romance phonetics and phonology. A machine-learning model, capable of automatically inferring graphical syllabification and stress assignment, along with the purpose of creating a data-base containing the quantitative and qualitative description of syllabification and stress in the Romance languages, could be the first important task in the greater challenge of tracing the similarities and differences between the Romance languages and, more important, between Romance and Latin. From a typological point of view, the study of syllabification and stress can shed a new light on the universal features that, by defining our phonoarticulatory and phonoacoustic apparatus, have guided the languages' development and change. Given the promising results of this analysis, the present study can establish the basis of a research of the syllable in other languages, either linguistically or typologically related to Italian.

One of the studies that address automatic syllabification in Italian belongs to Bigi and Petrone [16], who proposed a tool that performs rule-based automatic segmentation. Adsett and Marchand [17] and Adsett et al. [18] investigated whether data-driven approaches outperform rule-based approaches for a language with a low syllabic complexity, such as Italian. The authors reached the conclusion that even in this case data-driven systems are the more appropriate approach. In terms of machine learning, the tasks of automatically inferring syllable boundaries and predicting stress assignment can be naturally framed as sequence labeling problems. While automatic syllabification has received more attention recently [19, 20, 21, 22, 23, 24], stress placement has not been investigated as much [25].

Given the complexity of syllable applications and word syllabification, the presence of electronic resources dedicated to them becomes a necessity. While native speakers of a language generally do not have great difficulty in spelling words, the same cannot be said of those who learn a foreign language who often tend to apply their own rules to foreign words, and problems arise in automatic syllabification. This is because the rules of syllabification are linguistic rules, and they cannot always be easily modeled by the computer when there are no other linguistic factors that those rules take into account. For example, a rule that is present in many languages distinguishes between a vowel and a semivowel, but the computer is not able to easily recognize when the same sign has the value of a vowel and when it is a semivowel. Because of this, rule-based adaptations of syllabification systems [26] generally have higher errors, and many languages do not have an automatic syllabification system yet (for example, in the Python library, only a few languages have syllabification). The last few decades have brought the first data-driven syllabification systems.

However, in order to build such a system, training data is needed, and there are many cases in which the available data do not cover the whole language, and thus the systems have different results when the test corpus is changed.

Starting with these remarks, our main contributions are:

- We propose ItGraSyll (Italian graphical syllables), a dataset of $114,503$ Italian words, in orthographic form, containing annotations for their orthographic syllabification and stress placement[1]
- We perform quantitative and qualitative analyses of the previously built dataset.
- We analyze stress placement in the context of the Italian syllables.
- We propose an automatic system of syllabification for Italian words.

2. Quantitative Analysis

In this section we perform various measurements regarding the syllables and stress placement of Italian written words and analyze the results. We perform, on Italian, an investigation similar to a previous investigations conducted on Romanian by Dinu and Dinu [27], Dinu and Dinu [28].

2.1. Data

We build a dataset of Italian words starting from the online version of *Dizionario italiano De Mauro*,[2] which provides information regarding graphical syllabification and stress placement for the Italian vocabulary. Stressed syllables are also shown by having accents on the dominant vowel. Going further, this dataset will be referred to as ItGraSyll.

We performed several pre-processing steps. We cleaned the resulted dataset by removing duplicates, prefixes and suffixes in order to remain with the base word;

[1]The dataset is available for research purposes upon request at: https://nlp.unibuc.ro/resources.html#itgrasyll

[2]https://dizionario.internazionale.it/

abbreviations and unwanted punctuation marks such as dots, commas, apostrophes and dashes were also excluded so we can correctly process each word and its syllable division. Finally, the dataset consists of $114,503$ words in orthographic form having between one and eleven syllables. The distribution of words per number of syllables is represented in Table 1.

#syll.	#words	Examples
1	722	ai
2	5,960	àc-cia
3	23,286	àb-ba-co
4	41,253	a-ba-chì-sta
5	28,357	a-bi-tà-co-lo
6	10,829	ac-cu-mu-la-zió-ne
7	3,294	au-ten-ti-fi-ca-zió-ne
8	650	a-e-ro-mo-del-lì-sti-co
9	132	bi-o-me-te-o-ro-lo-gì-a
10	16	in-tel-let-tu-a-li-sti-ca-mén-te
11	5	ge-ne-ra-ti-vo-tra-sfor-ma-zio-nà-le

Table 1
Number of words per number of syllables.

Index	Syllable	Frequency
1	to	23943
2	re	18199
3	ta	12796
4	te	10987
5	si	10026
6	a	9142
7	co	8874
8	ri	8868
9	ca	8478
10	ra	8388
11	na	8367
12	ti	8184
13	ne	8112
14	men	7841
15	la	7175
16	di	6663
17	le	6555
18	li	6176
19	no	5748
20	lo	5479

Table 2
Top 20 most frequent syllables.

2.2. Syllables

We identified $\#Type_{syl} = 3730$ (type syllables) in Italian. The total number of syllables (token syllables) is $\#Token_{syl} = 483,931$. So, the average length of a word measured in syllables is $Words_{av-syl}$ = 483,931/114,503 = 4.226. The 114,503 words are formed of $\#Letters$ = 1,133,515 letters (graphemes). So, the average length of a word measured in letters is $Word_{av-let}$ = 1,133,515/114,503 = 9.899.

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: $LSyl_{token}$ = 1,133,515/483,931 = 2.342 b) the type syllables are formed of $\#TypeSyl_{let}$ = 13,576 letters. Thus, the average length of a type syllable measured in letters is $LSyl_{type}$ = 13,576/3,730 = 3.639.

These statistics are computed for the words extracted from the dictionary, which were considered to be equally weighted. This excludes any information relating to the frequency of the words with respect to writing or speech. For future research, large corpora of Italian texts can be leveraged in order to recompute these values and include frequency-based weights.

A list of the most frequent 20 syllables is included in Table 2.

2.3. Syllable Structure

We identified a total of 67 different consonant-vowel structures. The most frequent 7 structures cover almost 97% of the total. Depending on the type-token ratio, the most frequent consonant-vowel structures are the following: a) for the type syllables: cvc (25%), ccvc (20.9%), cvvc (7.79%). b) for the token syllables: cv (58%), cvc (15%), ccv (7%), cvv (4.74%) and v (4.32%). Moreover, we observe that the cv structure corresponds to 40 out of the most frequent 50 syllables from the dataset.

2.4. Stress Placement

We identified a total of 2,883 stressed syllables (type syllables). So, 847 syllables are never stressed. The most frequent 20 stressed syllables are represented in Table 3. We observe that the most frequent stressed syllable (*men*) has a very high stress ratio (90%) when we compare the stressed occurrences with all its occurrences (stressed and unstressed) in our database. While in the top 20 of all syllables, *men* is the only syllable of length 3 (on the 14th position), for stressed syllables there are a couple of other syllables with a length greater than 2 (*zio* on position 6 with 34% stress ratio, *gia* on position 19 with 65% stress ratio).

We investigate stress placement with regard to syllable structure and we provide in Table 4 the percentages of words having the stress placed on different positions (for top 5), counting syllables from the beginning and from the end of the words as well. We observe that in most cases the stress is placed on the second to last syllable.

Index	Syllable	Frequency	Stress ratio (%)
1	men	7120	90
2	ta	5809	45
3	na	3348	40
4	to	3254	15
5	la	2978	41
6	zio	2916	76
7	ti	2820	34
8	ca	2461	29
9	ra	2297	27
10	li	2239	36
11	ri	2100	24
12	tu	2024	62
13	za	2022	42
14	ni	1734	40
15	tri	1458	60
16	ma	1209	25
17	si	1144	11
18	da	1109	43
19	gia	1081	65
20	mi	1052	25

Table 3
Top 20 most frequent stressed syllables. The stress ratio indicates how often out of all the occurrences of the syllable in the corpus it appears as stressed.

Syllable	%words
1st	8,611
2nd	25,544
3rd	40,568
4th	25,593
5th	9,243

(a) counting syllables from the beginning of the word

Syllable	%words
1st	3,330
2nd	94,225
3rd	16,113
4th	14
5th	1

(b) counting syllables from the end of the word

Table 4
Stress placement for Italian.

2.5. Syllables' Usage

The syllables have a less intuitive behaviour, usually a small number of syllables cover a large part from a language. This is valuable for a large category of natural languages, including English, Dutch, Romanian [28], Korean, Chinese, etc. We investigate here if this empirical law is also applicable to Italian. We made this investigation both on stressed and general syllables.

2.5.1. General Syllables

The most frequent 30 Italian syllables (when stress placement is disregarded) cover almost 50% of $\#Token_{syl}$, the most frequent 100 cover 74% and the most frequent 150 syllables (i.e. 4% of $\#Type_{syl}$) cover 80% of $\#Token_{syl}$. Over this number, the percentage of coverage rises slowly. 2,281 (61%) syllables of type syllables occur less then 10 times, and 1,174 syllables occur only once (*hapax legomena*).

2.5.2. Stressed Syllables

A similar trend can be observed also for the stressed syllables. Further, we notice that the most frequent syllables cover a wide ratio of the total syllable frequency. For example, the 10 most frequent stressed syllable represent 31% of the total of stressed syllables, the top 50 syllables, 60% and the top 200 syllables, 81% of the token syllables. The values are plotted in Figure 1, for all syllables and for stressed syllables.

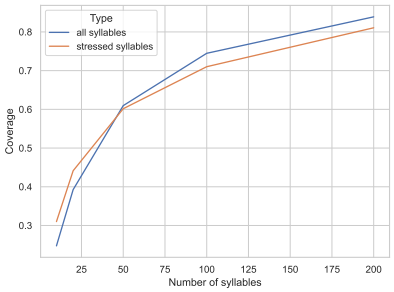


Figure 1: The coverage of most frequent syllables.

This results proves that the law is true for Italian too, a very small number of syllables cover a large part from Italian language (there are necessary only 150 syllables to cover 80% from language).

3. Minimum Effort Laws

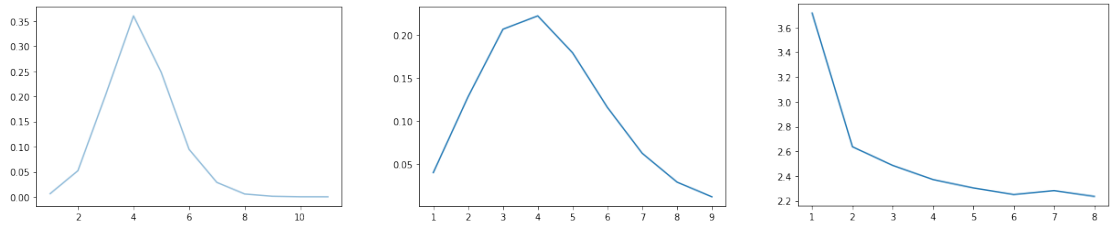
In this section we discuss two minimum effort laws that have been previously investigated for other languages and verify whether they apply for Italian as well.

3.1. Chebanow

Denoting by $F(n)$ the frequency of a word having n syllables and by $i = \sum nF(n)/\sum F(n)$ the average length (measured in syllables) of the words, Chebanow [29] proposed the following law between the average i and the probability of occurrences $P(n)$ of the words having n syllables:

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

For Italian, $i = 4.226$.



(a) The probability distribution of the length of words.

(b) Theoretical representation of the probability distribution of the length of words.

(c) Menzerath's Law: The more syllables in a word, the smaller its syllables.

Figure 2: Minimum effort laws.

Model	Hyphen Acc.	Hyphen F1	Word Acc.
GRU for syllabification w/o stress markers	99.74%	99.69%	97.61%
GRU for syllabification w/ stress markers	99.82%	99.79%	98.41%
GRU for stress prediction	—	—	94.45%

Table 5
Performance metrics computed for the automatic syllabification and stress prediction on the test set. We computed accuracy and F1 scores on the sequence labelling predictions for syllabification, in order to assess how well the model predicts the positions where the syllables split. Word level metrics were computed for both syllabification and stress prediction; this kind of metrics are more strict since any misplaced hyphen in the syllabification makes the entire prediction wrong.

In Figures 2a and 2b we plot the probability distribution of the length of words (in syllables) – the practical and theoretical representations.

We observe that the two curves have comparable shapes, with a more prominent peak for the probability distribution in Figure 2a; this peak can be influenced by the fact that it is determined based on all the words in the dictionary, where many 4-syllable words are present.

3.2. Menzerath

Menzerath's law – later generalized by the Menzerath-Altmann law [30] – states that the bigger the number of syllables in a word, the lesser the number of phonemes composing these syllables. In other words, Menzerath's law expresses a negative correlation between the length of a word in syllables and the lengths in phonemes of its constitutive syllables. In cognitive economy terms, this means that the more complex a linguistic construct, the smaller its constituents. The law is expressed as follows:

$$y = \alpha x^{\beta} e^{-\gamma x} \tag{2}$$

where y is the syllable length (the size of the constituent), x is the number of syllables per word (the size of the linguistic construct), and α, β, γ are empirical parameters. Figure 2c shows that the law is satisfied for Italian.

4. Automatic Syllabification and Stress Assignment

We further investigate how a deep-learning model can automatically infer the syllabification and stress assignment of Italian words, given their orthographic representation.

4.1. Methodology

Both tasks can be defined in terms of a sequence labelling problem, strategy which was previously successful used for Romanian[31, 32]. Let us consider, for example, the word *medaglione* (the Italian translation of the word "locket"). For syllabification we can label each letter from the word either with the label 1, denoting that a syllable starts from that letter, or with the label 0, meaning the respective letter is not the first letter in its syllable. Similarly, for identifying the stressed vowel, we can label its position with a 1 and all other letters are assigned the label 0. We thus obtain for our example the sequence 1010100010 for syllabification and the sequence 0000000100 for stress prediction (i.e. *me-daglió-ne*, the *o* vowel is stressed).

With these definitions, we can now construct machine learning models for labelling the character sequences. The model we propose is a recurrent neural network based on Gated Recurrent Units (GRU) [33]. The model architecture is comprised from the following components:

- a character embedding layer, producing 64-dimensional vectors for each unique character
- a stacked bidirectional GRU, with 3 layers and a 128-dimensional hidden state; a 0.2-rate dropout applied after each of the first two layers
- 0.5-rate dropout, after the last GRU layer, along with one-dimensional batch normalization
- a time-distributed fully-connected layer with 256 output nodes and ReLU activation
- a linear layer that projects the 256-dimensional vector into a single number, on which sigmoid activation is applied to infer the binary labels.

For training the models for both tasks, the dataset of words is split into 50% training examples and 50% test examples, unseen during training.

The loss function computed for the prediction made for a word, regardless of the task on which the model is trained, is the average of two terms: the first one is the average character-wise binary cross-entropy, while the second one is the root mean squared error computed between the vector of predicted labels and the ground-truth vector. The model is optimized using the Adam optimizer [34], with a learning rate of 0.0003, no weight decay, bath size of 32, and a LR scheduler that halves it every 5 epochs. The models are trained for 10-15 epochs.

For the task of automatic syllabification, we wanted to check if the presence of the stress markers affects the performance of the model. Because of that, we trained two models: the first one was trained using the spelling of the words with the stress markers removed, while the second one was trained with them included.

Stress Assignment Errors	
True	**Predicted**
bàlano	balanò
fèmore	femòre
dòlmen	dolmèn
tùtolo	tutòlo
pudìco	pùdico
corsìa	còrsia
Syllabification Errors	
True	**Predicted**
mu-o-ne	muo-ne
bion-da	bi-on-da
cli-en-te	clien-te
co-di-a-to	co-dia-to
ma-nu-brio	ma-nu-bri-o
spa-tria-to	spa-tri-a-to

Table 6
Examples of erroneous test predictions provided by the deep-learning models.

4.2. Results Anaysis

Table 5 contains the metrics computed on the test set, using the models trained for syllabification (both with and without stress markers) and the model trained for predicting the stressed vowel. We obtained a remarkable hyphen accuracy of 99.74% for syllabification without the stress markers, and, when we add the stress markers, we obtained an increasing accuracy, obtaining 99.82%. Including the stress markers into the data used for syllabification improved the metrics across the board, most notably with a $\sim$ 1% increase in word-level accuracy, which considering the large amount of data, and the high accuracy scores is a significant improvement (460 fewer syllabification mistakes as opposed to the approach that excludes stress markers). Regarding the stress prediction, we obtained an accuracy of 94.45%. Table 6 showcases a series of wrong predictions generated by the models on the tests sets for stress assignment and syllabification.

We also look into the accuracy scores computed for the test set, when it is bucketed based on the real number of syllables of the test words. These results are shown in Figure 3 and Table 7. For stress assignment, accuracy decreases to a global minimum for disyllabic words, then starts to increase again with the number of syllables. For the syllabification task, including the stress markers seems to outperform excluding them in most scenarios, while both accuracies achieve a peak around the 5 syllables mark. This result seems to align with the distribution of syllables in the dataset, i.e. obtaining higher scores for the number of syllables with more examples. For stress assignment errors, we also investigate the placement of the predicted stressed syllable in relation with the true one (see Table 8). 95.6% of the errors misplaced the stressed syllable at most one position to the left, or to the right, while almost two thirds of the erroneous predictions placed the stress on the first syllable to the right of the correct one.

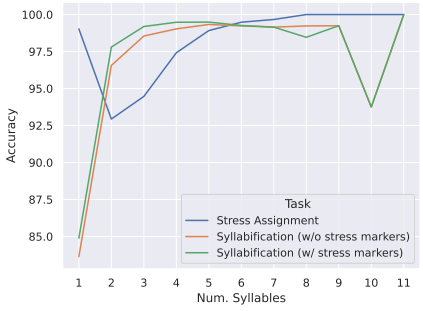


Figure 3: The test accuracies for each of the three tasks, computed independently on the test words, bucketed by their true number of syllables.

Num. Syllables	Num. Words	Stress Assignment	Syllabification (w/o SM)	Syllabification (w/ SM)
1	721	99.03%	83.63%	84.88%
2	5,960	92.94%	96.56%	97.80%
3	23,286	94.46%	98.55%	99.19%
4	41,253	97.42%	99.03%	99.48%
5	28,357	98.92%	99.33%	99.49%
6	10,829	99.48%	99.23%	99.26%
7	3,294	99.67%	99.15%	99.15%
8	650	100.0%	99.23%	98.46%
9	132	100.0%	99.24%	99.24%
10	16	100.0%	93.75%	93.75%
11	5	100.0%	100.0%	100.0%

Table 7
Similar to Figure 3 this table contains the actual values of the test accuracies for the three tasks: stress assignment, and syllabification with/without stress markers (SM) included. These scores are computed separately for words with the same number of syllables.

Stressed Syllable Delta	Num. Errors	Pct. Errors
-2	21	0.74%
-1	804	28.38%
0	95	3.35%
1	**1,809**	**63.85%**
2	102	3.60%
3	2	0.07%

Table 8
Starting from the incorrect predictions for stress assignment, we compute how far the assigned stress is from the actual one, in numbers of syllables (delta). A delta of -2 means that the predicted stressed syllable is the second one to the left of the correct stressed syllable. A delta of 0 in this situation means that the algorithm predicted the stressed vowel incorrectly, but the prediction sits inside the correct stressed syllable.

5. Conclusions

In this paper we have investigated graphical syllabification and graphical stress assignment for Italian words. We have started by building ItGraSyll, a dataset of Italian graphical syllabified words, with stress annotations as well, on which we have performed several quantitative and qualitative analyses, including the verification of two minimum effort laws for the case of Italian. Finally, we have proposed a recurrent neural network machine learning model for automatic syllabification and stress assignment for Italian written words. For stress prediction we have obtained 94.45% word-level accuracy, and for syllabification we have obtained 98.41% word-level accuracy and 99.82% hyphen-level accuracy. In future work we intend to extend the analysis from dictionary level to corpus level and to investigate other languages as well.

Acknowledgments

We want to thank the reviewers for their useful suggestions. Research supported by the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, SiRoLa project, number PN-IV-P1-PCE-2023-1701, Romania.

References

[1] S. Suyanto, Incorporating syllabification points into a model of grapheme-to-phoneme conversion, International Journal of Speech Technology 22 (2019) 459–470.

[2] V. N. Vitale, F. Cutugno, A. Origlia, G. Coro, Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique, Neural Comput. Appl. 36 (2024) 6875–6901. URL: https://doi.org/10.1007/s00521-024-09435-1. doi:10.1007/S00521-024-09435-1.

[3] A. Dinu, L. P. Dinu, On the syllabic similarities of romance languages, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings, volume 3406 of *Lecture Notes*

in Computer Science, Springer, 2005, pp. 785–788. URL: https://doi.org/10.1007/978-3-540-30586-6_88. doi:10.1007/978-3-540-30586-6_88.

[4] J. P. Kincaid, L. R. P. F. Jr., R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel, Research Branch Report, Millington, TN: Chief of Naval Training, 1975.

[5] G. Marco, J. de la Rosa, J. Gonzalo, S. Ros, E. González-Blanco, Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches, IEEE Access 9 (2021) 51734–51746.

[6] A. M. Ciobanu, L. P. Dinu, On the romanian rhyme detection, in: Proceedings of COLING 2012: Demonstration Papers, 2012, pp. 87–94.

[7] L. Hjelmslev, The syllable as a structural unit, in: the Proceedings of the 3rd International Congress of Phonetic Sciences (Ghent), 1938, volume 266, 1938.

[8] F. De Saussure, Course in general linguistics, Columbia University Press, 2011.

[9] D. Russo, The Notion of Syllable across History, Theories and Analysis, Cambridge Scholars Publishing, 2016.

[10] M. Loporcaro, Syllable, segment and prosody, in: The Cambridge history of the Romance languages, 2011, pp. 50–108.

[11] W. Meyer-Lübke, Grammaire des langues romanes, volume 4, H. Welter, 1906.

[12] M.-D. Glessgen, Linguistique romane: domaines et méthodes en linguistique française et romane, Armand Colin, 2007.

[13] F. S. Miret, Fonética histórica, in: Manual de lingüística románica, Ariel España, 2007, pp. 227–250.

[14] F. d'Ovidio, W. Meyer-Lübke, Grammatica storica della lingua e dei dialetti italiani, volume 368, U. Hoepli, 1906.

[15] G. Rohlfs, T. Franceschi, Grammatica storica della lingua italiana e dei suoi dialetti: Morfologia, (No Title) (1968).

[16] B. Bigi, C. Petrone, A generic tool for the automatic syllabification of italian, A generic tool for the automatic syllabification of Italian (2014) 73–77.

[17] C. R. Adsett, Y. Marchand, Are Rule-based Syllabification Methods Adequate for Languages with Low Syllabic Complexity? The Case of Italian, in: P. Wagner, J. Abresch, S. Breuer, W. Hess (Eds.), Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007, ISCA, 2007, pp. 58–63.

[18] C. R. Adsett, Y. Marchand, V. Keselj, Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of italian, Comput. Speech Lang. 23 (2009) 444–463. URL: https://doi.org/10.1016/j.csl.2009.02.004. doi:10.1016/j.csl.2009.02.004.

[19] K. A. Rogova, K. Demuynck, D. V. Compernolle, Automatic syllabification using segmental conditional random fields, in: Computational Linguistics in the Netherlands Journal, volume 3, 2013, pp. 34–48.

[20] L. P. Dinu, V. Niculae, O. Sulea, Romanian syllabication using machine learning, in: I. Habernal, V. Matousek (Eds.), Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings, volume 8082 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 450–456.

[21] J. Krantz, M. W. Dulin, P. D. Palma, Language-Agnostic Syllabification with Neural Sequence Labeling, 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (2019) 804–810.

[22] V. N. Vitale, L. Schettino, F. Cutugno, On incrementing interpretability of machine learning models from the foundations: A study on syllabic speech units, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023, volume 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3596/paper51.pdf.

[23] O. Sulea, L. P. Dinu, B. Dumitru, Full inflection learning using deep neural networks, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing - 19th International Conference, CICLing 2018, Hanoi, Vietnam, March 18-24, 2018, Revised Selected Papers, Part I, volume 13396 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 408–415. URL: https://doi.org/10.1007/978-3-031-23793-5_33. doi:10.1007/978-3-031-23793-5_33.

[24] M. Petrillo, F. Cutugno, A syllable segmentation algorithm for english and italian., in: INTERSPEECH 2003, 2003, pp. 2913–2916.

[25] Q. Dou, S. Bergsma, S. Jiampojamarn, G. Kondrak, A Ranking Approach to Stress Prediction for Letter-to-Phoneme Conversion, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1, ACL '09, Association for Computational Linguistics, 2009, p. 118–126.

[26] L. P. Dinu, An approach to syllables via some extensions of marcus contextual grammars, Grammars 6 (2003) 1–12. URL: https://doi.org/10.1023/A:1024089129146. doi:10.1023/A:1024089129146.

[27] L. P. Dinu, A. Dinu, On the data base of romanian syllables and some of its quantitative and cryp-

tographic aspects, in: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006, European Language Resources Association (ELRA), 2006, pp. 1795–1798.

[28] L. P. Dinu, A. Dinu, On the behavior of romanian syllables related to minimum effort laws, in: Proceedings Workshop Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages, co-located with RANLP 2009, Borovets, Bulgaria 2006, 2009, pp. 9–13.

[29] S. Chebanow, 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) 99–102.

[30] G. Altmann, Prolegomena to Menzerath's Law, Glottometrika 2 (1980) 1–10.

[31] A. M. Ciobanu, A. Dinu, L. P. Dinu, Predicting romanian stress assignment, in: G. Bouma, Y. Parmentier (Eds.), Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden, The Association for Computer Linguistics, 2014, pp. 64–68. URL: https://doi.org/10.3115/v1/e14-4013. doi:10.3115/V1/E14-4013.

[32] L. P. Dinu, A. M. Ciobanu, I. Chitoran, V. Niculae, Using a machine learning model to assess the complexity of stress systems, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014, European Language Resources Association (ELRA), 2014, pp. 331–336. URL: http://www.lrec-conf.org/proceedings/lrec2014/summaries/1200.html.

[33] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[34] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).