A Rule-based Approach to Generating Large Phonetic Databases for Romanian

Results of the AFLR Project

Ștefan - Stelian Diaconescu,
Monica - Mihaela Rizea, Mihaela Ionescu,
Andrei Mincă, Monica Rădulescu
Research & Development Department
SOFTWIN

***Abstract*— This paper presents a rule-based approach for generating a large phonetic database for Romanian. The knowledge base is developed by means of the GRAALAN (Grammar Abstract Language) system. By inspecting dictionaries and corpora, we generate a phonetic database over 100,000 lemmas. Our database has a high degree of accuracy ensured by our rule-based method applied for generating phonetic transcriptions.**

Keywords—phonetics; linguistics; natural language processing; GRAALAN; Romanian language

I. INTRODUCTION

One of the aims of the AFLR[1] Project is to conduct research on Romanian phonetics in order to develop complex phonetic knowledge bases, including rules for automatic phonetic transcriptions. This research would be facilitated by the already existing instruments represented by the GRAALAN metalanguage (Grammar Abstract Language) [1], which was especially defined for describing linguistic knowledge, and by the already developed software tools. A large phonetic database will represent an important resource for an Automatic Speech Recognition System, which is also one of the aims of AFLR.

The current state of research in the domain of Romanian phonetics and phonology underlines the necessity of a number of applied studies with the aim of obtaining (formalized) rules for automatic phonetic, as well as the necessity of large machine-readable or even printed dictionaries to record phonetic transcriptions for Romanian words, using an internationally approved standard. A number of Romanian dictionaries such as [2], [3] or [4] can also be considered important theoretical guides for word pronunciation and syllable division/stress patterns. [3], for example, indicates pronunciation using a number of signs that imply a "compromise" between the "scientific transcription" and current Romanian "orthography", in order to serve the purpose of accessibility to a large public (see [5]). However, there is currently no Romanian dictionary providing phonetic transcriptions, syllable boundaries, and stress patterns for a large number of words, using a computer-readable internationally approved standard.

[1] AFLR – Analiza Fonetică a Limbii Române ("Romanian Language Phonetic Analysis")

II. BRIEF DESCRIPTION OF THE GRAMMAR ABSTRACT LANGUAGE (GRAALAN)

The GRAALAN metalanguage is a very powerful instrument allowing the formalization of multiple linguistic sections: Alphabet, Syllabification, Morphology, Inflection Rules, Syntax, and Inflection Forms. Therefore, the GRAALAN metalanguage is used for describing:

- Alphabets (including various types of information, such as representation of phonemes and graphemes, or even correspondences between groups of letters and phonemes, and so on);
- Phonetic representations of lemmas (at the lexicon level);
- The inflection rules in both normal and phonetic alphabet;
- Syllabification rules (for normal and phonetic alphabet), and, implicitly, for obtaining the syllables of the language;
- The synthetic and analytic inflected forms (accompanied by their phonetic transcription and syllabification).

The GRAALAN Alphabet section is characterized by a series of features, generally applicable to all natural languages that are being represented in this metalanguage, but also by language-specific elements. The general features imply the use of two standards: Unicode 7.0 and the International Phonetic Alphabet (IPA) 2005. The Alphabet section of a language is generally defined by the following subsections:

- **Phonetic Alphabet** (IPA characters), which is generally represented as a list of phonemes and contains internal characters, which are language-specific, but also external characters (i.e., non-specific characters that are mainly used in the phonetic transcription of unadapted neologisms).
- **Normal Alphabet**, which refers to a list of graphemes (letters) that are specific to a certain language (i.e., internal characters). Similarly to the Phonetic Alphabet, the Normal Alphabet permits the description

978-1-5090-6497-7/17/$31.00 ©2017 IEEE

of non-specific, external characters when the lexicon of the language that is being described contains foreign spellings or when we specify an etymon spelled with characters of the original language.

- **Special Characters** can be internal or external (see, for example, quotation marks that have specific representations for some languages); they include punctuation symbols (that can have special connector/separator functions attached), currency symbols, diacritical marks, etc.
- **Stress Characters**, which defines the suprasegmental characters for marking primary or secondary stress patterns.
- **Groups** define the correspondence between letters and phonetic characters (in other words, in this section we represent a number of typical contexts for the grapheme/phoneme correspondence). The phonetic representation includes (primary/secondary) stress variants for the groups that contain vowels.
- **Classes** contain lists of labels that are defined in the previously mentioned Alphabet subsections and are used in order to optimize GRAALAN-encoded rules (for example, the Syllabification section can contain rules that call for classes already defined in the Alphabet section of a language).

It is important to mention that defining the phonetic alphabet of Romanian using the compliable language GRAALAN implied, first of all, solving a number of theoretical issues, in the absence of previous normative studies that reflected the position of the Romanian Academy, and that standardized the inventory of the IPA symbols considered specific for representing the phonemes of the Romanian language.

Using the GRAALAN metalanguage, we have previously described, in a series of other projects, a phonetic database covering a lexicon that contained, at that moment, more than 75,000 lemmas (which was already a large database). We should specify that our phonetic transcriptions use the IPA standard (adapted to Romanian) for the representation of phonemes, syllable boundaries, and stress patterns (the IPA transcription is rendered with the standard SIL font). We have generally used the principle of broad phonetic transcription – see [20] (which meets a degree of generalization that corresponds to our purpose of extending the description to other languages as well). Our method of generating the phonetic database is largely described in former studies such as [6].

During the AFLR project, we have mainly focused on providing phonetic transcriptions, syllabifications, and stress patterns for a series of neologisms – mostly not adapted to the Romanian spelling/morphology – representing common language, but also specialized vocabularies. These were either present in the Romanian dictionaries (see the list of dictionaries in the following section) without pronunciation suggestions, or were generated after investigating several corpora, internally created or provided by one of the project partners, "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy (details in the following sections).

The GRAALAN encoding of our phonetic knowledge bases made them easily adaptable to the project requirements: new characters, groups, and classes had to be defined in the Alphabet section, and new syllabification rules had to be defined in order to permit the description of this challenging vocabulary, represented in the Romanian contemporary language.

Another advantage of GRAALAN language is represented by the high degree of accuracy ensured by our rule-based method applied for generating phonetic transcriptions. We know that other methods and approaches in the field of automatic grapheme to phoneme conversions (such as machine learning systems based on artificial neural networks or hybrid systems) generally consider the guarantee of a high level of accuracy a difficult task (see, for example, the results reported in [7]). The low degree of accuracy is also very problematic in the case of languages with alphabetic (or non-phonemic) orthography, such as English. In our approach, the semi-automatic method applied by making use of GRAALAN predefined groups and internally developed applications[1] ensures a high degree of accuracy since the databases are filled exclusively by linguistic experts. The high level of accuracy is also obtained due to our strategies of verification at the end of each development stage.

III. THE EVOLUTION OF THE ROMANIAN PHONETIC DATABASES IN THE AFLR PROJECT

A. *Starting point*

At the beginning of this project, the Romanian language database described using the GRAALAN metalanguage had the following dimensions:

- 76,837 lemmas (with phonetic transcriptions) and 116,074 meanings
- 12,119,349 inflected forms[2] (i.e, 861,718 synthetic forms and 11,257,631 analytical forms) with phonetic transcriptions
- 13,207 multiword expressions (MWEs) (with phonetic transcriptions) and 14,052 meanings
- 90,393 semantic relations between meanings

The linguistic knowledge base (LKB) was developed during the previous projects and it was based on the collection of words present in the dictionaries in use at that moment [8]. Starting from this LKB, the paper entitled *Fonetica Limbii Române* (*Romanian Language Phonetics*) was created ([9], [10], [11], [12]). It is an interdisciplinary study written in 4 volumes, that proposes a number of solutions both for theoretical, purely linguistic problems and for issues related to the field of Natural Language Processing (NLP), such as: creating applied studies, large databases and, not least,

[1] For a detailed description of these applications, see [6].
[2] The process of generation of the inflected forms from the lexicon lemmas is largely described in [6].

applications with a high level of performance that exploit these resources for Romanian.

The current stage of the Romanian language has necessitated updates to this LKB, in particular by introducing neologisms, which are more and more frequent in the current language. Thus, the first step in creating a Large Phonetic Database was selecting a large number of lemmas missing from the current LKB. The sources taking into consideration for this process were digital corpora, digital dictionaries, and standardized infrastructural linguistic knowledge bases. The list of selected words contains 24,719 potential words, which increased the LKB by over 25%.

As expected, most of these potential words, i.e. 21,028 words, are neologisms. It should be mentioned that all the selected words were reviewed by linguists from "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy (ILIR), partner in this project, so that the pieces of information introduced in the LKB to be accurate.

B. Working procedure

The selected words were introduced in the LKB by the members of the ILIR partner, using software instruments developed by SOFTWIN during previous projects. The newly added data often implied linguistic decisions regarding the phonetic transcription, the syllabification or the definitions.

During the process of introducing new words, updates were necessary, both of the tools and of the rules described in GRAALAN. Digital online dictionaries were also inspected. Depending on their specific, the following 7 digital dictionaries were considered: DEX 98 [8], DEX 09 [13], DEX 12 [14], NODEX (2002) [15], MDA (2002) [16], MDN (2000) [17], DSL (1997) [18]. In 4 of them, DEX 98, DEX 09, DEX 12, and MDN (2000) we identified words not yet covered by the GRAALAN knowledge bases, as can be seen from the Table 1.

One of the digital corpora inspected was ZiareRom, offered by the ILIR partner. This led to the identification of 66,503 word candidates that were not yet covered by the GRAALAN knowledge bases; however, most of them proved to be abbreviations, words with spelling mistakes or words written without diacritical marks, which motivated our decision of removing them from the list of new words to be added to the lexicon.

Other digital corpora that we used were SOFTWIN digital library (consisting of electronic books from different fields) and the European Union documents made freely available on the official website. By comparing, by means of a searching system, the words from the corpora with the words from the LKB, we obtained a list of 238,000 words unknown by the system. 75% of these words were proper names (of persons, institutions, products, etc.). In these circumstances, the detection of the "new" unregistered words was quite laborious. About 40,000 words from the list generated by the searching system were inspected, but about 1,200 words were selected to be introduced in the LKB.

The process of introducing words implied filling several fields with different types of linguistic information:

- description of the lemma in alphabetic format
- description of the lemma in phonetic format (IPA)
- selection of the inflection situation (the part of speech and its morphological features: gender, number, conjugation etc.)
- selection of the inflection rule (requirements of the declination/conjugation type)
- alphabetic and euphonic syllabification
- the meanings and the usages (popular, regionalism etc.)
- description of the supplements (the main inflected forms) in terms of alphabetical, phonological, morphological features, as well as their alphabetic and euphonic syllabification; For example, in the case of a noun, the supplement is its plural undefined Nominative form and for a verb, the supplements consist of the form of the present indicative, first person singular and the masculine participle.
- information of the lemma structure, such as the infinitive particle, the reflexive pronoun, and the short form of the infinitive

All the words introduced have been reviewed in order to ensure that there is no piece of information missing on the phonetic, syllabification, morphological or semantic level.

C. Final point

At the end of this project, the Romanian language database described using the GRAALAN metalanguage had the following dimensions:

- 100,708 lemmas[3] (with phonetic transcriptions) and 142,299 meanings
- 13,219 MWEs (with phonetic transcriptions) and 14,066 meanings
- 184,075 semantic relations between meanings

We should mention that, in the current stage, the priority was the introduction of (single) words, not of expressions/fixed phrases. Even if the description of MWEs is covered by the GRAALAN metalanguage and some tests have been made in this direction, introducing MWEs will be the subject of another project.

TABLE I. IDENTIFIED WORDS NOT YET COVERED BY THE GRAALAN KNOWLEDGE BASES

Dictionary	Number of words
DEX 98	187
MDN	21,028
DEX 09	934
DEX 12	2
multiple sources	1,386

[3] The inflected forms (along with their phonetic transcriptions) for the newly introduced lexicon lemmas are currently in process of being generated.

Thus, at the end of AFLR, we obtained a phonetic database over a lexicon with more than 100,000 lemmas, a database with a high degree of accuracy ensured by our rule-based method applied for generating phonetic transcriptions.

IV. CHALLENGES IN GENERATING LARGE PHONETIC DATABASES FOR ROMANIAN

Building a phonetic database over a large lexicon (comprising more than 100,000 words) implied overcoming some theoretical issues related to the contemporary Romanian language, but also some practical issues of data management and digitization. The GRAALAN system ensures an efficient data structure and possibilities for describing numerous linguistic aspects. Therefore, from the point of view of structuring and storing the necessary information, the GRAALAN system provided all the necessary conditions. Updating the Alphabet subsection from the system was most often the right solution to solve the linguistic problems that occurred during this project.

At the beginning of the project, the Alphabet subsection contained the data given in Table II.

During the process of building the phonetic database, we encountered several types of challenges:

A. *The lack of normal characters in the Alphabet subsection, necessary for foreing words*

The import of foreign words in Romanian language involved introducing words with foreign spellings in the Romanian dictionaries. In order to introduce these words in our LKB, an update of the subsection Alphabet was needed, in terms of adding the missing characters.

Here are some examples of external characters that had to be introduced in the Normal Alphabet section. All new characters had to be included in special groups in order to facilitate the phonetic transcription. The groups that contained vocalic sounds implied three types of representations: with primary/secondary stress and versions with no stress (this ensures rendering the correct stress pattern when data is introduced by linguistic experts):

/* å */ example of word: landsmål [lˈandsmol], bokmål [bˈukmol]

```
character
code = "å"
type = external
label = a_over_ring
order key = 2.160

/*Å */
character
code = "Å"
type = external
label = A_over_ring
order key = 2.159
```

TABLE II. ALPHABET SUBSECTION DATA AT THE BEGINNING OF AFLR PROJECT

Data type	Number
Normal characters	74
Phonetic characters	36
Special characters	78
Stress characters	2
Groups	635
Classes	24

Groups associated with new characters:

```
/*å, Å in the word landsmål */
  group
  code = (("å","Å")
  [("&mid_back_rounded;")])
  label = o_from_a_over_ring

/*å, Å */
   group
   code = (("å","Å")
   [("&primary_stress;&mid_back_rounded;")])
   label = o_from_a_over_ring_primary_stress

/*å, Å */
   group
   code = (("å","Å")
   [("&secondary_stress;&mid_back_rounded;")])
      label = o_from_a_over_ring_secondary_stress
```

B. *Impossible phonetic transcription for imported foreign words*

Many of the foreign words recently added to Romanian dictionaries or frequently used in spoken and written language are using the spelling and pronunciation of the language of origin. This means that, in order to facilitate the phonetic transcription, we had to bring to date the phonetic rules by including new phonetic characters and new groups. During the project, we have covered all the detected cases.

During this project, three phonetic characters and 152 groups were added.

Here are examples of external characters that had to be introduced in the Phonetic Alphabet section.

/* ɔ̃ */ example of word: fondue [fɔ̃dy]

```
character
code = "&#x0254;&#x0303;"
type = external
label = nasalised_open_mid_back_rounded
stressed = yes
order key = 1.158
```

/* ɑ̃ */ example of word: champleve [ʃɑ̃ləvˈe]

```
character
code = "&#x0251;&#x0303;"
type = external
label = nasalised_open_back_unrounded
```

```
        stressed = yes
        order key = 1.155

    /* θ */ example of word: commonwealth [kˈomənwelθ]
        character
        code = "&#x03B8;"
        type = internal
        label = voiceless_dental_fricative
        order key = 1.184

    Groups associated with new characters:
    /* on  in the word fondue */
        group
        code = (("oO"/"nN")
        [("&nasalised_open_mid_back_rounded;")])
        label = nasalized_o_from_on

    /* on */
        group
        code =  (("oO"/"nN")

[("&primary_stress;&nasalised_open_mid_back_rounded;")])
        label = nasalized_o_from_on_primary_stress

    /* on */
        group
        code = (("oO"/"nN")

[("&secondary_stress;&nasalised_open_mid_back_rounded;")
])
        label = nasalized_o_from_on_secondary_stress

    /* amp in the word champlevé */
        group
        code = (("aA"/"mM"/"pP")
        [("&nasalised_open_back_unrounded;")])
        label = nasalized_a_from_amp

    /* amp */
        group
        code = (("aA"/"mM"/"pP")

[("&primary_stress;&nasalised_open_back_unrounded;")])
        label = nasalized_a_from_amp_primary_stress

    /*amp    */
        group
        code = (("aA"/"mM"/"pP")

("&secondary_stress;&nasalised_open_back_unrounded;")])
        label = nasalized_a_amp_secondary_stress

    /* e in the word champleve*/
         group
         code = (("e","E")
         ["&close_mid_front_rounded;"])
         label = o_with_diaresis_from_e
```

TABLE III. SOME EXAMPLES OF WORDS THAT NEEDED NEW GROUPS IN THE ALPHABET SUBSECTION, IN ORDER TO FACILITATE THE PHONETIC TRANSCRIPTION.

Word	Phonetic transcription	New groups added in order to facilitate the phonetic transcription
abbevilian	abvilˈan	bbe [b]
barbecue	bˈarbəku	ue [y]
baud	bo	aud [o]
bidonville	bidonvˈil	lle [l]
board	bˈord	oa [o]
boogie	bˈugi	gie [ɟi]
bordeaux	bordˈo	eaux [o]
byte	bˈajt	te [t]
cappuccino	kapuʧˈino	cci [ʧˈi]
chalet	ʃalˈet	et [e]
copywriter	kˈopirajtər	wr [r]
gourde	gˈurd	de [d]
landsmål	lˈandsmol	å [o]
mall	mˈol	a [o]
montagnard	montagnˈard	gna [nja]
muguet	mygwˈet	guet [ɟe]
neokeynesism	neocejnesˈizm	key [ce]
offline	ˈofflajn	ne [n]
rinascimento	rinaʃimˈento	sc [ʃ]
saintlucian	səntluʧiˈan	ai [ə]
stracchino	stracˈino	cch [c]
ukulele	jukejlˈejli	u [ju] ku [cej] e [ej] e [i]
yeoman	jˈowmen	eo [ow]

```
    /* e */
        group
        code = (("e","E")
        [("&primary_stress;&close_mid_front_rounded;")])
        label = o_with_diaresis_from_e_primary_stress

    /* e */
         group
         code = (("e","E")
         [("&secondary_stress;&close_mid_front_rounded;")])
         label = o_with_diaresis_from_e_secondary_stress

    /* th in the word commonwealth */
         group
         code = (("tT"/"hH")
         [("&voiceless_dental_fricative;")])
         label = theta_from_th
```

At the end of this project, the Alphabet subsection contains the data given in Table IV.

C. *Collecting and building the linguistic information associated with corpora selected words*

The process of adding corpora-selected words to the linguistic database implied collecting and building all the linguistic information requested by the GRAALAN system. This presupposed verifying the contexts of each word in order to detect the meanings and to build the inflection forms. In addition, the words already existing in the database received new meanings, currently available in the language.

TABLE IV. ALPHABET SUBSECTION DATA AT THE END OF AFLR PROJECT

Data type	Number
Normal characters	76
Phonetic characters	39
Special characters	78
Stress characters	2
Groups	787
Classes	24

V. CONCLUSIONS

In this paper we presented the development of one of the objectives of AFLR project: generating a large phonetic database for Romanian language.

Using the GRAALAN metalanguage we succeeded to build a phonetic database with more than 100,000 lemma type entries, covering the relevant linguistic information: phonetic transcriptions, syllabifications, glosses, morphology, semantic relations, inflection. The high degree of accuracy is ensured by our rule-based method applied for generating phonetic transcriptions.

This large phonetic database represents an important linguistic resource for Romanian language and it will be used in speech recognition algorithms and other natural language processing systems.

VI. ACKNOWLEDGMENT

This work has been realized through the Partnerships Program in priority areas - PNII, developed with the support of MEN - UEFISCDI, project no. PN-II-PT-PCCA-2013-4-1451, contract no. 332/2014.

VII. REFERENCES

[1] S. Diaconescu, "GRAALAN – Grammar Abstract Language Basics", GESTS International Transaction on Computer Science and Engineering, Vol.10, No.1, 2005.

[2] I. Vintilă-Rădulescu, "Dicţionar normativ al limbii române ortografic, ortoepic, morfologic şi practic", Bucharest: Corint, 2009.

[3] I. Vintilă-Rădulescu, "Dicţionarul ortografic, ortoepic şi morfologic al limbii române", ediţia a II-a, revăzută şi adăugită, Bucharest: Editura Univers Enciclopedic, 2005.

[4] F. Şuteu and E. Şoşa, "Dicţionar ortografic al limbii române", Bucharest: Vestala/Atos, 1993.

[5] "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy, "Dicţionarul ortografic, ortoepic şi morfologic al limbii române", second edition, Bucharest: Univers Enciclopedic Gold, 2010.

[6] S. Diaconescu, M. Rizea, F. Codîrlaşu, M. Ionescu, M. Rădulescu, A. Mincă, Ş. Fulea, "Methods for Automatic Generation of GRAALAN-based Phonetic Databases", Proceedings of the 8th Speech Technology and Human-Computer Dialogue Conference SpeD, 135-142 (2015).

[7] J. Domokos, O. Buza, and G. Toderean, "Romanian phonetic transcription dictionary for speeding up language technology development," Language Resources and Evaluation, vol. 49, pp. 311-325, Jun 2015.

[8] "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy , "Dicţionarul explicativ al limbii române", second edition, Bucharest: Univers Enciclopedic, 1998.

[9] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, "Fonetica limbii române", (Vol. I, Elementele metalimbajului GRAALAN), ISBN 978-1514314784, 2015.

[10] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, "Fonetica limbii române", (Vol. II, Dicţionarul morfologic şi fonetic al limbii române, A-L), ISBN 978-1514315125, 2015.

[11] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, "Fonetica Limbii Române", (Vol. III, Dicţionarul morfologic şi fonetic al limbii române, M-Z), ISBN 978-1514315262, 2015.

[12] Ş. S. Diaconescu, F. C. Codîrlaşu, M. Ionescu, M. M. Rizea, M. Rădulescu, A. Mincă, Ş. Fulea, "Fonetica Limbii Române", (Vol. IV, Dicţionarul fonetic al silabelor limbii române şi dicţionarul fonetic de rime al limbii române), ISBN 978-1514315422, 2015.

[13] "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy , "Dicţionarul explicativ al limbii române", second edition, Bucharest: Univers Enciclopedic Gold, 2009.

[14] "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy , "Dicţionarul explicativ al limbii române", second edition, Bucharest: Univers Enciclopedic Gold, 2012.

[15] "Noul dicţionar explicativ al limbii române", Editura Litera Internaţional, 2002.

[16] "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy , "Micul dicţionar academic", second edition, Bucharest: Univers Enciclopedic, 2002.

[17] Florin Marcu, "Marele dicţionar de neologisme", Editura Saeculum, 2000.

[18] Angela Bidu-Vrănceanu, Cristina Călăraşu, Liliana Ionescu-Ruxăndoiu, Mihaela Mancaş, Gabriela Pană Dindelegan, "Dicţionar general de ştiinţe ale limbii", Bucharest: Editura Ştiinţifică, 1997.

[19] Peter Roach, "English Phonetics and Phonology", Cambridge: Cambridge University Press, third edition, 2000.

[20] D. Jones, P. Roach, J. Hartman, J. Setter and (editors), CEPD 2003, "Cambridge English Pronouncing Dictionary" CD-ROM, Cambridge: Cambridge University Press, 2003.

[21] E. Vasiliu, "Scrierea limbii române în raport cu fonetica şi cu fonologia", Bucureşti: Universitatea din Bucureşti, 1979.

[22] E. Petrovici, "Sistemul fonematic al limbii române", Studii şi cercetări lingvistice 7, 1956.

[23] I. Chiţoran, "The Phonology of Romanian: A Constraint-Based Approach", Berlin; New York, Mouton de Gruyter, 2002.

[24] I. Maddieson, "Phonetic Cues to Syllabification" in Phonetic linguistics: Essays in honor of Peter Ladefoged, Orlando, Florida, Academic Press, Inc., 1985.