On European Portuguese Automatic Syllabification

Catarina Oliveira[1]*, Lurdes Castro Moutinho*[1]*, Antonio Teixeira*[2]

[1]Center for Language and Cultures, [2]Dep. Electronics and Telecommunications/IEETA
University of Aveiro, Aveiro, Portugal

coliveira@dlc.ua.pt lmoutinho@dlc.ua.pt ajst@det.ua.pt

Abstract

This paper presents three methods for dividing European Portuguese (EP) words into syllables, two of them handling graphemes as input, the other processing phone sequences. All three try to incorporate linguistic knowledge about EP syllable structure, but in different degrees. Experimental results showed, for the best method, percentage of correctly recognized syllable boundaries above 99.5 %, and comparable word accuracy. The much simpler finite state transducer based method also achieved a good performance, making it suitable for applications more interested in speed and memory footprint. Being syllabification an essential component of many speech and language processing systems, proposed methods can be useful to researchers working with the EP language.

1. Introduction

In classic generative phonology theory the syllable is not the basis for the application of phonological rules. For the later generative multilinear models, the syllable is an important linguistic unit, hierarchically organized. In this theoretical framework, various literature contributes support what is called the syllabic structure model "Onset-Rhyme".

Syllabification in TTS conversion is important for two reasons[1]. First, it helps the implementation of certain letter-to-phoneme rules. Second, syllabification is essential in enhancing the quality of synthetic speech since detecting the syllable will help in modelling duration and improve the synthesized speech intonation. Syllabification is also useful in Automatic Speech Recognition (ASR). A first effort in using syllables for European Portuguese (EP) ASR was [2].

The usefulness of the syllabification information for the formulation of certain phonological rules, combined with the generally reported syllabic base of EP hyphenation, motivated our explorations in the development of automatic syllabification methods having graphemes as input. The present work follows the line of linguistic research arguing that linguistic theories should be implementable and tested.

Automatic syllabification was approached recently using different theories and methods, such as: optimality theory (eg. [3]), and the automatic acquisition of syllable grammars [4].

The paper is structured as follows: The next two sections summarize knowledge about EP syllable and present, briefly, an algorithm for syllabification proposed by two leading EP Phonology researchers; Section 4 presents two methods we developed for syllabification, one of them being available in two versions, one for orthography input, the other for syllabification of phones; Sections 5 and 6 present evaluation methodology and respective results; The final section presents the conclusions.

2. European Portuguese Syllables

The concept of syllable based on the binary branching model with rhyme has proved to be productive and efficient in the European Portuguese description [5, 6]. According to this perspective "syllable is a multidimensional object with an internal structure that has a hierarchical organization where the onset and the rhyme constitute a branching structure" [5]. Like this, a syllable branches in onset and rhyme and this last one branches in Nucleus and Coda. Each syllabic constituent is associated to one or, in the maximum, two skeletal positions, represented by X in the syllabic tree.

The Onset can be simple (not branching) and be connected to a skeletal position (e.g. *pé* "foot"), or branching (complex) if it is connected to two skeletal positions (e.g. *prato* "dish"). Any PE consonant can occupy this position, but it's possible that segmental material isn't connected to the onset (empty onset - *árvore* "tree"). Regarding complex onsets, the consonantic groups constituted by stop+liquid and by fricative+liquid respect the sonority principle and dissimilarity condition, even though the first sequences are much more frequent [7]. Other consonantic groups (e.g. pn - *pneu* "tire"; ps - *psicologia* "psychology"; pt - *raptar* "to kidnap"; ft - afta "aftae"; mn - *amnésia* "amnesia") clearly violate the universal principles. To explain these apparent violations, Mateus and d'Andrade [5] propose the existence of subjacent empty nucleus.

The same way the syllabic rhyme constituent may be simple (*bola* "ball"); or branching, when it is made up by the Nucleus and Coda (e.g. *porta* "door"). The nucleus may be associated to a minimum of one and the maximum of two positions in skeleton level. In EP, the syllabic nucleuses are always vowels. Like this, the nucleus can be filled by any vowel (non-branching nucleus - *casa* "house") or by a diphthong (branching nucleus - *pai* "father"), formed by a vowel followed by a glide, which is also included in the nucleus. The rising diphthongs (GV) create difficulties regarding the determination of the syllabic limit. The GV sequence alternates, in the production with the VV structure ([piad6] / [pjad6] - "joke"), which means that the semi-vowel of the diphthong has in the base a phonological vowel and syllabic boundary exists between the two elements of the diphthong ($V.V$). The problem does not place itself in the level of the base (phonologic) syllabification, but in the resyllabification, which results in the loss of the syllabic feature and semi-vocalization of the firs diphthong element with the consequent join of the two syllables (GV). The phonetic alternation GV/VV is not verified, in some contexts which involve a semi-vowel [w] preceded by velar stop [k] or [g]. In this case, we can consider that: 1) the semi-vowel is part of a branching onset; 2) the semi-vowel belongs to a branching nucleus; 3) the structure is associated to a non-branching onset, being the sequence velar consonant+glide a monophonematic unit, i.e. a labialized velar stop ([k^w], [g^w]). Another case of ambiguity respects structures VGV (e.g. *areia* “sand”).

Regarding Coda, only three consonants, /l,s,r/, with its different realizations, can occupy this position [6]. In previous analysis it is admitted, although, the occurrence of complex Codas. Barbeiro (1986) [8] considers the sequences /rs/,/ls/,/ns/, being the last one the most frequent.

Mateus (1993) [9] gives some examples of syllables with complex Codas (*abstrair* “to abstract”, *perspectiva* “perspective”), emphasizing the fact these sequences always present the form C/s/ and may happen, mainly in pre-stressed syllables, with non-branching nucleus. Afterwards, the author suggests another explanation for these structures, postulating the existence of empty nucleus between consonants considered in previous analyses as Coda. The consonant /s/ is also the only one able to follow a complex nucleus (*claustro*), increasing to three the number of segments associated to the rhyme.

Statistic studies about the EP syllabic structure carried out in the 1990’s [7, 10] permit us to conclude that the CV structure is, by far, the most frequent in polysyllables as in monosyllables. The percentages confirm the tendency of EP language to the syllabic opening, greatly certified by various authors.

3. Mateus and d’Andrade Algorithm for EP Syllabification

The syllabic division results from the application of mechanisms which try to give syllabic roles (rhyme level) to the segments (segment level).

Mateus [6], and later Mateus and d’Andrade [5], in the context of the “Onset - Rhyme” syllable theory, proposed a rule based syllabification algorithm, considered more adequate to syllabic structure creation in EP than the template-matching approach.

According to the adopted theory each syllabic constituent is associated to at least one skeletal position.

In the adopted approach, traditionally called “all nuclei first approach”, we start with the construction of rhymes, in agreement with the languages restrictions. Like this, the first rule applied regards the nucleus, to which all X **[+ syllabic]** are associated. Seeing that the semi-vowels of the falling diphthong are also a part of the nucleus, the second part of the rule includes the rest of the X **[-consonantic]** in the adjacent nucleus to the left. With the creation of the nucleus, automatically a rhyme is constructed.

The second rule implies an association of each X **[+consonantic]** preceding a nucleus to an onset. A sequence of two consonants is inserted in the same onset, as long as it respects the sonority principle and the dissimilarity condition. The second part of the same rule allows to associate the remaining **[+consonantic]** /s,l,r/ , which are Coda.

The presence of extra-syllabic consonants to the left of the onset (X **[+ cons]**) not associated to any constituent and not integrated in the syllabic structure results in the creation of empty positions. Then, an empty nucleus to the left of the onset is created (e.g. *pneu* - pV.neu), with the corresponding position in the skeleton.

A re-application of the second rule allows the association of the consonants without position in the skeleton to an onset, now placed before an empty nucleus. In the case of a rhyme not being preceded by an onset, the creation of this last and the correspondent position in the skeleton line is demanded.

We assume that the syllable in the EP is obligatorily constituted by an onset and by a rhyme, although one of these constituents may be empty or have no phonetic realization.

Finally, the underlying representations are divided in syllables.

The authors have varied, along the years, the order of rules application, particularly regarding coda formation. We tried to follow their initial proposals.

4. Syllabification Methods

In this section are presented, briefly, two different automatic syllabification methods we developed based on the previous mentioned theory on the EP syllable and mainly on the algorithm proposal of Mateus and d’Andrade, from now M&A, (section 3). The first method, using finite state transducers (FSTs), uses essentially the general description of the syllable constituents. The second follows closely the M&A algorithm.

4.1. Finite state orthography based syllabification

This process was developed as part of a grapheme-phone conversion (g2p) system for EP [11], based on manual rules, implemented by FSTs. Syllabification was the first step in the conversion, being syllable boundaries used in (some of) the grapheme-phone rules definition. Acting as an auxiliary step, syllabification was not separately evaluated until now.

Syllabification was implemented following the proposal of Bouma [12] for Dutch, consisting in the composition of 3 transducers in sequence. First marks the nucleus, next inserts syllable boundaries, the last removes nucleus marks. Before this 3 step process proposed by Bouma, we have a transducer that includes a mark representing an empty nucleus, an “V”, between a list of not allowed consonant sequences. Example: *pneu* “tire” is transformed by this first transducer in *pVneu.*

For the first transducer, we created a list of consonants and consonant clusters allowed in the onset position. As mentioned, all consonants can be onsets, as well as stop+liquid and fricative+liquid sequences.

Also the list of possible nucleus was defined. In this definition we took in consideration that: all vowels, stressed or not, can fill a nucleus; followed by a glide vowels form falling diphthongs, and both integrate nucleus; nasal diphthongs, so frequent in EP, can also be nucleus. Resulting from the empty nucleus insertion, the “V” mark can also be nucleus. An “N” is used as a nasalization mark, being introduced by the system in a pre-treatment of the word, before the syllabification process. In EP the graphemes “n” and “m” represent nasality when placed before a consonant. In our system, we only considered the falling diphthongs.

Definition of coda list followed closely the description of what can constitute this syllable part, presented in section 2, restricted to non-branching codas.

Defined all lists of allowable graphemes for the three syllabic constituents, the transducers act in sequence, based on the command `replace`:

1. Nucleus marking, inserting an “@” before and after the possible nucleus: `replace([[]:@,id(nucleus),[]:@,[],[])`
2. Syllable boundaries marking, based on the generally accepted description of EP syllable structure: rhyme with obligatory nucleus and optional onset and coda. Optionality is signalled by “^”.
   `replace([]:'|', [@,coda^],[onset^,@])`

3. Finally, removing of nucleus delimitation marks: `replace(#:[], [], []).`

4.2. Implementation of Mateus and d'Andrade (M&A) algorithm

The input is plain text. The output of the method and of all the several steps is structured in an XML document. XML provides a powerful, flexible and intuitive way to structure data. XML processing is done in Perl using XML::DOM implementing the XML Document Object Model (DOM).

We implemented two versions of the algorithm: one handling orthography as input, the other capable of processing phone sequences. Being the second one a more direct implementation of the proposed algorithm, we only present here information regarding the grapheme based algorithm.

Several steps, corresponding roughly to the M&A algorithm, implemented, for now, as separate programs, allow careful inspection of the results of each algorithm step.

Having as input a list of words, a preliminary step creates a node in the hierarchy for each word.

A second preliminary step processes each word, splitting it in graphemes and creating a sub-node for each. This sub-node is a potential syllable. To each potential syllable, in addition to the grapheme, several attributes are added regarding stress (using [13] code) and consonantic and syllabic features. It is also in this step that grapheme groups are detected; being treated similarly to individual graphemes (e.g. double "r" is grouped and associated to only one potential syllable).

After the two preliminary steps, the first step of M&A algorithm is performed, assigning the value Nucleus to the PartSyllable attribute of **[+syllabic]** nodes. This is performed in the two sub-steps proposed by the M&A algorithm.

The second step of the algorithm creates onsets. In our implementation, first, the **[+consonantic]** before a nucleus are marked as being Onset in the PartSyllable attribute, second, the **[+consonantic]** before the recently created Onset are evaluated regarding the sonority principle. If the two have a difference of sonority high enough [7], increasing toward the nucleus, this second node is marked as onset and associated to the syllable of the other onset.

Third step processes remaining [+consonantic] marking the ones allowed to occupy coda position with the PartSyllable value of Coda.

In the definition of both the table of sonority values, for second step, and the list of allowed codas, for the third step, we processed a list of around 100 k words with these steps, monitoring the graphemes processed and decisions made. With this information, the table and list were corrected and completed.

The final step deals with empty nucleus creation and finalization of the algorithm.

4.2.1. Example

As an example the XML fragment generated for word *reserva* "reserve" is (showing only the relevant information):

```
<WORD SPELLING="reserva"  STRESS="rese1rva">
<syllable>
<letter Stress="-" ... PartSyllable="O">r</letter>
<letter Stress="-" ... PartSyllable="N">e</letter>
</syllable>
<syllable>
<letter Stress="-" ... PartSyllable="O">s</letter>
<letter Stress="1" ... PartSyllable="N">e</letter>
<letter Stress="-" ... PartSyllable="C">r</letter>
</syllable>
<syllable>
<letter Stress="-" ... PartSyllable="O">v</letter>
<letter Stress="-" ... PartSyllable="N">a</letter>
</syllable>
</WORD>
```

and the automatically generated graphic representation:

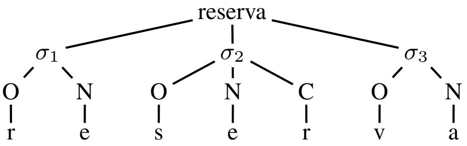

Figure 1: Word=reserva, stress=rese1rva

5. Evaluation Methodology

The evaluation of the two developed algorithms and a publicly available was performed in terms of the number of agreements in the syllable boundaries and the number of non-existent boundaries inserted, both expressed as percentages. First, is the percentage of correct decisions, defined as the ratio between the number of boundaries considered correctly placed and the total number of existent syllable boundaries; second, the percentage of insertions, defined as the ratio between the number of incorrectly inserted boundaries and the total number of boundaries. An overall figure of performance, called precision, can be obtained by subtracting the second from the first. Also, the number of correctly processed words (word accuracy) was calculated. For now, only syllable boundaries were evaluated, leaving a more detailed evaluation of each of the syllable constituents for later. These evaluation parameters were obtained automatically, using a specific developed program in Perl. The problematic words were manually analyzed and classified in terms of the error type, giving further insights in the methods weaknesses.

Besides the development corpus (of 1006 words), for evaluation, and due to the non-public availability of a corpus having syllable boundaries and pronunciation for EP, we created two corpus. First, consists in 2076 common words, corresponding to a fraction of the so-called "Português Fundamental" (Fundamental Portuguese) [14]. Second, consists of 1303 words randomly selected from the Público corpus created by the Portuguese Project Linguateca [15] from the newspaper Público editions. This list of words contains words of longer length and higher complexity.

Phonetic pronunciation, using SAMPA phonetic alphabet, was added to the corpora. To make the process faster, first an automatic grapheme-phone system was used, being the result manually corrected. Also, at this stage, pronunciation was aligned to the orthography to enable automatic comparisons, making explicit to what graphemes correspond no pronunciation and relations between two graphemes and one phoneme. Later, and also manually, syllable boundaries were added to the orthography. Automatically, these boundaries were transposed to the pronunciation.

6. Results

Our evaluation included our methods and, as term of comparison, a publicly available EP syllabification algorithm [13]. All methods were evaluated in the two test corpora.

The results of three evaluation metrics (percentage of correct syllable boundaries, percentage of insertions and word accuracy) for the 4 methods tested are presented in Table 1. The results are presented separately for the two corpora.

		M&A	FSTs	PT::PLN	M&A
Corpus		Graph.	(Graph.)	(Graph.)	Phones
Test1	%Cor	99.77	99.27	96.62	98.46
	Ins	0.08	1.17	0.19	0.26
	%WAc	99.57	98.80	93.93	97.88
Test2	Correct	99.59	98.92	96.06	98.80
	Ins	0.15	0.75	0.15	0.03
	%WAc	98.85	97.40	88.95	96.47

Table 1: Results (in percentage) of the evaluation of the various syllabification methods against two manual segmented corpus. Table presents both the percentage of correctly positioned syllable boundaries (%Cor), the percentage of insertions (%Ins), and percentage of correctly syllabified words (Word Accuracy, represented by %WAc).

Clearly the extension of Mateus e d'Andrade algorithm to process the words orthography attains a very good performance. Not only in terms of correctly detected syllable boundaries, but also by the high number of correctly syllabified words and low percentage of insertions. The phone-based version has a good but lower performance. Decrease affects particularly Word Accuracy. The FST method has also a good performance, suffering from a higher level of insertions. The public algorithm had a somewhat lower performance.

Looking in more detail to the words incorrectly syllabified by each method: M&A based on graphemes several errors are due to error in stress assignment, prefixation, incorrect insertion of empty nucleus; most of the FST errors come from stress, different approaches to handling sequences such as "ct" and "qu" and "gu", errors in identification of a nasal vowel as nucleus; for M&A phone based errors come mainly from alignment problems; the public method errors come in the majority from the use of an hyphenation approach in sequences such as "cç" (divided between adjacent syllables even when the first is not realized phonetically), acceptation of raising diphthongs (not considered by our methods and corpus annotation), acceptation of complex syllable onsets in clear violation of the sonority principle, and, worst, failures in some other aspects (e.g. nasal vowel syllabification, coda segments wrongly integrated in onset of the next syllable, separation of digraphs such as "qu").

Direct comparison between the results of all the three grapheme based syllabification methods and comparison between the two versions of the M&A algorithm implementation was also performed. The results are not included due to space limitations.

7. Conclusion

In this paper we presented two methods for EP automatic syllabification. One used a finite state transducers approach. The other consists in the implementation of an adapted version of Mateus and d'Andrade syllabification algorithm proposal.

An evaluation of the performance of the developed methods results in a clear demonstration that the extension to grapheme input of the originally proposed algorithm for base syllabification was very successful. Also, the two variants of the algorithm compared favourably with the less linguistic approaches of the FST and publicly available methods. Nevertheless, due to the not so much inferior performance of FST method and simplicity, it is an interesting approach for less linguistic interests.

Future work will address the lexical stress assignment and morphologic analysis of words, particularly to handle syllabification errors in prefixes. Being the work presented motivated essentially by our efforts in the development of a complete text-to-speech system, incorporating the most recent knowledge and linguistic theories, the new grapheme based syllabification method will be used to provide information to the g2p module, in phase of re-implementation. We intend to use the same approach, based on DOM and use of XML, in this new version of the g2p module.

8. Acknowledgements

The first author thanks Universidade de Aveiro and FCT (Portuguese Research Agency) the PhD scholarships supporting her work in Linguistic Processing for TTS systems.

9. References

[1] Y. A. El-Imam, "Phonetization of arabic: rules and algorithms," *Computer Speech & Language*, vol. 18, pp. 339–373, October 2004.

[2] H. Meinedo, "Utilization of syllabic information in the recognition of continuous speech," Master thesis, Univ. Técnica de Lisboa/IST, 2000, in Portuguese.

[3] M. Hammond, "Syllable parsing in english and french," 1995.

[4] K. Mueller, "Probabilistic syllable modeling using unsupervised and supervised learning methods," PhD thesis, University of Stuttgart, Institute of Natural Language Processing(IMS), Stuttgart, 2002.

[5] M. H. M. Mateus and E. d'Andrade, *Phonology of Portuguese*. OUP, 2000.

[6] M. H. M. Mateus, "A silabificação de base em português," in *X Encontro Assoc Portug. Ling. (APL)*, Évora, 1994.

[7] M. Vigário and I. Falé, "A sílaba no português fundamental: uma descrição e algumas considerações de ordem teórica," in *IX Encontro Assoc Portug. Ling. (APL)*, 1993.

[8] Barbeiro, "Estrutura silábica do português. o papel da sílaba na análise dos processos fonológicos e fonéticos," Master Thesis, Univ. Lisboa, 1986.

[9] M. H. M. Mateus, "Onset of portuguese syllables and rising diphtongs," in *Proc. of the Workshop on Phonology*, Coimbra, 1993.

[10] E. d'Andrade and M. do Céu Viana, "Sinérese, diérese e estrutura silábica," in *IX Encontro Assoc Portug. Ling. (APL)*, 1993.

[11] C. Oliveira, S. Paiva, L. de Castro Moutinho, and A. Teixeira, "Um novo sistema de conversação grafema-fone para o Português Europeu baseado em transdutores," in *II Congresso Internacional de Fonética e Fonologia*, 2004.

[12] G. Bouma, "Finite states methods for hyphenation," *Journal of Natural Language Engineering*, no. 1, 2002.

[13] J. J. Almeida, A. Simões, and P. Rocha, "Lingua-PT-PLN-0.06," 2003. [Online]. Available: http://www.cpan.org

[14] F. Nascimento, L. Marques, and L. Segura, "Português fundamental: Métodos e documentos," INIC-CLUL, Lisboa, Tech. Rep., 1987.

[15] [Online]. Available: http://www.linguateca.pt