INVESTIGATING NEW SYLLABLE PROTOTYPES FOR THE PORTUGUESE LANGUAGE

Sara Candeias[a] *& Fernando Perdigão*[a,b]

[a]Polo de Coimbra, Instituto de Telecomunicações, Coimbra, Portugal;
[b]DEEC, Polo II, Universidade de Coimbra, Coimbra, Portugal
saracandeias@co.it.pt; fp@co.it.pt

ABSTRACT

This paper presents research results on new syllable structures for the Portuguese language. The location of the syllable boundaries is a well known problem of nonconsensual resolution in Portuguese, mainly when the acoustic-phonetic constraints are taken into account. Based on various acoustic-phonetic restrictions of Portuguese speech, new syllable splitting rules were studied in a corpus of 400K Portuguese words. All these words were automatically converted into sequences of C (consonants) and V (vowels) and also related to the new syllable structures. When syllable structures were outlined in a statistical form, some mapping confirmed our expectations, others apparently ran against the expectations and others appeared to be new. These results provide clues about Portuguese syllabification architecture and can also guide the development of improved complementary models for syllable processing in speech applications.

Keywords: syllable prototypes, acoustic-phonetic parameters, Portuguese speech

1. INTRODUCTION

There are various well-known studies on the Portuguese language about syllable structures (hereafter, PS – Portuguese syllables). Mateus and d'Andrade [7] discuss the internal structure of the syllable in European Portuguese; Freitas and Santos, among others, in [4] describe some phonotatic restrictions for the segment combi-nations into the syllable, while Vigário, Fronta and Martins in [5] and [16], for instance, explore the syllable as a prosodic unit which is hierarchically structured. Regarding the relevance of frequency information in linguistics, the works [1] and [18] (based on "*Português Fundamental*" corpora) and references [6] and [19] (based on samples from the "*Português Falado - Documentos Autênticos*"), among others, are worth mentioning. Although the subject of syllables in Portuguese has been studied for a long time, a clear acoustic-phonetic support for it has not been entirely developed. This is important in several areas, including speech synthesis and recognition where the syllable unit has been used to derive rules governing stress determination (e.g. [2, 13, 14]). Unlike the studies mentioned above which examined splitting PS, this paper's approach consists in presenting a complete set of all syllable structures for the Portuguese language and deals with acoustic-phonetic parameters closer to actual speech. Based on segment sequences of C (consonants) and V (vowels), syllable structures were automatically selected and manually checked in order to accumulate knowledge of PS prototypes.

This study has three main purposes which complement the previous approaches: (i) to argue that syllable prototypes dealing with acoustic-phonetic parameters are closer to actual speech (ii) to describe statistically the syllable structure prototypes in Portuguese using a large Portuguese corpus, and (iii) to introduce PS as an intermediary level between words and phones that can serve as a consistent linguistic tool to explore complementary acoustic models for speech processing in Portuguese.

Section 2 of this paper briefly states the syllable and the syllabification concepts. Section 3 describes the corpora and outlines the general methodology used to locate syllable boundaries. Section 4 presents the results of a statistical analysis of PS and Section 5 draws conclusions.

2. SYLLABLE AND SYLLABIFICATION

Theoretically, a syllable is a unit of organization for a sequence of speech sounds. Since vowels are normally more perceptible than consonants, each vowel of an utterance will correspond to a peak in the curve of perceptibility or audibility. However, this does not mean that certain sounds, according to this context, cannot function as the syllable peak, which is common for a vowel, or as the flanking unit of this peak, which is normal for a consonant.

The generativist description of Portuguese phonology, described in [8], has been used as a framework to describe the constitution of a syllable. The PS has been postulated as a speech unit of rhythmic organization, involving a sound group around a vowel and produced by a single expiratory movement (see [5] and [16], for instance). It is assumed that the rhyme is the mandatory nucleus, which corresponds to a vowel, even if the nucleus is empty.

Syllabification means the determination of the place of syllable boundaries in a word. This is not a consensual procedure. Sometimes syllabification deals more with a concept of "syllable" that corresponds to the written form, while in other situations it is correlated in some way with audibility. Much work on the so-called sonority sequencing principle introduced as early as the 19th century by Sievers [12] and more recently by Selkirk [11], has reflected the idea that phones may be arrayed in terms of power and relative audibility (a sonority scale). Clements' studies (e.g. [3]) are worth mentioning as well, since they suggest that sonority is related to the relative resonance of speech sounds. The sonority scale view was also adopted in some work on the Portuguese language, for example by Mateus and d'Andrade [7] and [8] p. 39-53 and Vigário and Falé [18] who showed that phones form syllables according to their perceived intensity, based on their condition of dissimilarity. A good example of the sonority sequencing principle in Portuguese language is the single syllable word <*graus*> (meaning, degrees). The first consonant in the syllable onset is <*g*>, which is a stop, the lowest on the sonority scale; next is <*r*>, a liquid which is more sonorous, then we have the vowel <*a*> [a] - the sonority peak; next, in the syllable coda, is <*u*>, the glide [*w*], and finally, a fricative, <*s*> [S]. By applying these principles any syllable is defined so that its center, canonically a vowel, constitutes a sonority peak that is preceded (onset) and/or followed (coda) by a sequence of segments (none or more consonants), with progressively decreasing sonority values. It also means that the sonority values of the syllable segments can be determined by a sonority hierarchy. A great deal of work (e.g. in [5, 16]) has been done in the last decade on syllabification, particularly by intonational and prosody experts, who have also shown some violations of these principles at the phonetic level. In fact, the location of the syllable boundaries is a problem of nonconsensual resolution, mainly when acoustic-phonetic restrictions are considered. It is also worth noting that the definition of the sonority principle and its relation to the sonority hierarchy is not enough to establish all possible PS sequences – we shall see later that discrepancies of this kind appear in some consonant clusters such as |p ɾ| or |dv| or |bs|.

To summarize, the method of syllabification depends mainly on the conventions of linguistic communities. As different languages exhibit different syllabic structures, there cannot be a universal theory of base syllabification. A descriptiton of the syllable structure prototypes deriving from acoustic-phonetic properties is required for the Portuguese Language. Complementing the existent approaches with a statistical form, the study described here takes a step in the above mentioned direction.

3. METHODOLOGY

3.1. Data

The corpus used in this study is a selection of Portuguese words from the *Natura* Project's corpora [10] that was developed for spelling purposes. This list has about 600k words including proper names, acronyms, abbreviations and common foreign-words. It also includes compound words and pronominal tenses. In order to represent the Portuguese language more accurately, acronyms and foreign-words were deleted from this corpus, resulting in a list of 401,345 Portuguese words. To begin with, the syllabification process was done automatically using the *Natura* Project tool (the Lingua::PT::PLN module). Subsequently, some syllable prototypes were changed, according to new syllabification rules, and some were manually corrected.

3.2. The analysis technique

One goal of this work is to map all sequential Portuguese C/V syllable patterns and relate them to syllabic structuring. According to our main idea (outlined below), in order to solve some syllable boundary problems we have to apply some new splitting rules.

The problem that happens most often is the syllable splitting of sequential vowels, because they are considered 'growing diphthongs' (i.e., adjacent heterosyllabic vowels, which are combined into diphthongs or even triphthongs). For example, the first syllable in the words <*reunir*> (to meet) and <*biunívoco*> (biunivocal)

or the last syllable in the word <*gasóleo*> (diesel oil) may be split using the tendency to admit a prevocalic vowel as a glide: *reu|nir, biu|n í|vo|co, ga|s ó|leo* (e.g. in [9]). Alternatively, as in the present study, the syllables may be split, taking into account either a tendency to make location syllable boundaries coincide with potential morphemes (for instance, in <*re+unir*> or in <*bi+un ívoco*>).

Another principle is the tendency to link a sequence of consonants in order to put plosives always in an onset location. Accordingly, with this last convention, words like <*advertir*> (to warn), <*obter*> (to obtain), <*opção*> (option) or <*absentismo*> (absenteeism) have their syllable boundary between the vowel and the plosive: <*a|dver|tir*>, <*o|pção*> and <*a|bsen|tis|mo*> (versus the classical syllable splitting marked as <*ad|ver|tir*>, <*op|ção*> and <*ab|sen|tis|mo*>).

Obviously, there are some ambiguous cases. The case around the diphthong <*iu*> is one of the most typical of them, particularly because <*i*> and <*u*> could both be a glide. There are some cases in which we decide to take <*iu*> as a diphthong (e.g. <*ciu|mar*> (to be jealous) or <*en|viu|var*> (to widow)), since it is possible to confront them with similar contexts in which the diphthong is undone (e.g. <*ci|úme*> (jealous) or <*vi|úvo*> (widower)). However, the same syllabification procedure cannot be applied to all other cases.

We are in a tricky field here and there is certainly no consensus on the answer to the question where the location syllable boundary is. We argue that our outlined syllable splitting is more consistent than canonical location syllable boundaries and this helps to strengthen our belief that our PS structure prototypes are powerful tools for Portuguese speech applications, in particular.

By applying all these criteria, partly automatic and partly manual, to the corpus of roughly 400K words, about 1.8M syllable items have been mapped (see Table 1). For each word, a CV syllable sequence is generated. For example, the word <*resultados*> (results) is split into 4 syllables: <*re|sul|ta|dos*>. The resulting syllables are then converted into a CV structure: CV|CVC|CV|CVC. Since this procedure has no grammatical cons-traints, any syllable can follow any CV order. However, the boundary configuration allowed only 26 different PS structure sequences, as shown in Table 1. The number of different instances of each prototype is indicated in the last column.

Table 1: Number(#) of occurrences of both the structures and the sequence prototypes of PS, as observed in the corpus. The percentage (%) of occurrences is also shown. *Closed syllables are shown in bold.

Structure of PS prototypes*	PS Occurrences (#)	%	PS Sequence prototypes (#)
CV	856974	48.38	230
CVC	370514	20.92	662
CCV	140671	7.94	302
V	140042	7.91	15
VC	93818	5.3	53
CVV	62117	3.51	204
CCVC	37161	2.1	408
CVVC	33544	1.89	158
VVC	9146	0.52	12
CCVVC	8368	0.47	77
CCVV	6378	0.36	154
VV	4622	0.26	12
CVCC	2095	0.12	48
VCC	1699	0.1	6
CVVV	1523	0.09	10
CCVCC	1465	0.08	7
CVVVC	671	0.04	3
CCCV	415	0.02	15
C	20	<0.02	2
CCCVV	20	<0.02	2
CCCVC	16	<0.02	6
VVCC	4	<0.02	1
CCCVVC	3	<0.02	3
CC	3	<0.02	2
CVCCC	2	<0.02	1
CVVCC	1	<0.02	1
26	*1771292*		*2394*

4. STATISTICAL ANALYSIS

As mentioned earlier, the corpus contains approximately 1.8M syllable occurrences, with about 2.4K distinct PS sequence prototypes (Table 1). Percentages for the occurrences of the different PS structure prototypes in the corpus are also given in the table.

Although there are more closed than open syllable prototypes (17 and 9, respectively), open syllables are significantly more common than closed syllables (68.47% and 31.53%, respectively). As mentioned before, there are 26 PS prototypes, but they rely mainly on two: |CV| and |CVC|, which account for 48.38% and 20.92% respectively of all occurrences. From the remainning PS, it should be noted that the |CC…| prototypes (such as |CCV|, |CCVV|, |CCVC|, |CCCV|, |CCVVC|, |CCVCC|, |CCCVC| and |CCCVVC|) account for 10.98% of the corpus. This particular structure is mainly due to the syllabification rule of plosive location. The plosives are merged with the onset of the following syllable to form a branching onset and increase the consonant cluster structures by around 2.48%. In

canonical Portuguese grammar, a consonant grapheme cannot constitute a syllable, but if an acoustic-phonetic constraint is applied, a syllable composed of one consonant in a final word position would be possible. In fact, we agree on consonant syllables in Latin words such as <*de*|*fi*|*ci*|*t*> or <*ha*|*bi*|*ta*|*t*>, and in the verbal conjugation of the verbs <*ter*> (to have) and <*vir*> (to come) (and their derivative verbs), expressly in the 3rd person plural, present tense). Some cases are as follows: <*abs*|*tẽ*|*m*> (they keep away from), <*a*|*dvẽ*|*m*> (they arrive), and <*tẽ*|*m*> (they have). The need for these 'partial' syllables (<|*m*|> and <|*t*|>) arises from the presence of specific function graphemes in the Portuguese language, which could be pronounced as an added V, mainly an [ə] or a [ɐ], sometimes evident in the acoustic signal. Notice that the schwa [ə], which may or may not be perceived (thus manipulating the number of syllables), is one of the most complex aspects of Portuguese phonology (see [15] so far). To explain the presence of an apparent syllable without any vowel, Mateus and d'Andrade [8] proposed the existence of a subjacent empty nucleus. Moreover, |C| PS structure is rare in the Portuguese language, accounting for <0.02% of all words in the corpus.

5. CONCLUSION

The potential of using statistical information methods to increase specification of Portuguese syllables is emphasized in this work. Supported on a corpus of about 400k Portuguese words, this study presents statistics of syllable prototypes, showing, for example that about 70% of all syllables have either the prototype |CV| or |CVC|. Although some evidence already existed, the exact distribution of these prototypes dealing with acoustic-phonetic parameters closer to actual speech has never been reported before, as far as we know. Some less frequent prototypes are problematic because there is no consensus on the syllabification rules, especially in VV and some CC clusters. The results presented here in a statistical form follow various syllabification principles. A full list of all PS may be obtained from the authors.

6. ACKNOWLEDGMENTS

Sara Candeias would like to thank the *Science and Technology Foundation-FCT* for the Post-PhD grant (SFRH/BPD/36584/2007).

7. REFERENCES

[1] Andrade, E., Viana, M.C. 1994. Sinérese, Diérese e Estrutura Silábica. *Actas do IX Enc. Nac. APL*. Lisboa: Colibri, 31-42.

[2] Candeias, S., Perdigão, F. 2008. Conversor de grafemas para fones baseado em regras para português. In Costa, L., Santos, D., Cardoso, N. (eds.), *Perspectivas sobre a Linguateca: 10 anos.* Lisboa: Linguateca, 99-104.

[3] Clements, G.N. 1990.The role of the sonority cycle in core syllabification. In Kingston, J., Beckman, M. (eds.), *Papers in Laboratory Phonology I*. Cambridge: Cambridge University Press. 283-333.

[4] Freitas, M.J., Santos, A.L. 2001. Contar (Histórias de) Sílabas. Descrição e Implicações para o Ensino do Português como Língua Materna. *Colibri Ed.*, Lisboa.

[5] Frota, S., et al. 2002. Language discrimination and rhythm classes: Evidence from Portuguese. *Proc Speech Prosody 2002* Aix-en-Provence, 315-318.

[6] Frota, S., et al. 2006. FreP: An electronic tool for extracting frequency information of phonological units from Portuguese written text. *Proc. Language Resources and Evaluation* Genoa, 2224-2229.

[7] Mateus, M.H., d'Andrade, E. 1998. The syllable stucture in European Portuguese. *DELTA [online]* 14(1), 13-32.

[8] Mateus, M.H., d'Andrade, E. 2000. *The Phonology of Portuguese.* Oxford: Oxford University Press.

[9] Portal da Lingua Portuguesa. *http://www.portaldalinguaportuguesa.org/*

[10] Projecto Natura. *http://natura.di.uminho.pt/wiki/doku.php?id=projectonatura*

[11] Selkirk, E.O. 1984. Phonology and Syntax: The Relation between Sound and Structure. MIT Press.

[12] Sievers, E. 1881. *Grundzüge der Phonetik*. Leipzig: Breitkopf und Hartel.

[13] Teixeira, A., et al. 2006. On the use of machine learning and syllable information in European Portuguese grapheme-phone conversion. In Vieira, R., et al (eds.), *Computacional Processing of the Portuguese Language, PROPOR 2006* Springer.

[14] Teixeira, J.P., et al. 2000. Divisão Silábica Automática do Texto Escrito e Falado. *Proc. PROPOR'2000* Atibaia, SP, Brasil.

[15] Veloso, J. 2007. Schwa in European Portuguese: The Phonological status of [ɨ]. In Crouzet, O., Angoujard, J.-P. (eds.), *Actes des JEL'2007, Schwa(s)*, 55-60.

[16] Vigário, M., et al. 2003. From signal to grammar: Rhythm and the acquisition of syllable structure. *Proc. of the 27th Annual Boston University Conf. on Language Development*. Dommerville. Mass: Cascadilla Press, 809-821.

[17] Vigário, M., et al. 2005. Frequências no Português: A ferramenta FreP. *Actas do XX Enc. Nac. APL*. Lisboa: Colibri, 897-908.

[18] Vigário, M., Falé, I. 1994. A Sílaba no Português fundamental: Uma descrição e algumas considerações de ordem teórica. *Actas do IX Enc. Nac. APL*. Lisboa: Colibri, 465-478.

[19] Vigário, M., Martins, F., Frota, S. 2006. A ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português. *Actas do XXI Enc. Nac. APL*. Lisboa: Colibri, 675-687.