10.12871/CLICIT2014115

A generic tool for the automatic syllabification of Italian

Brigitte Bigi, Caterina Petrone
Laboratoire Parole et Langage, CNRS, Aix Marseille Université
5 avenue Pasteur, 13100 Aix-en-Provence, France
{brigitte.bigi,caterina.petrone}@lpl-aix.fr

Abstract

English. This paper presents a rule-based automatic syllabification for Italian. Differently from previously proposed syllabifiers, our approach is more user-friendly since the Python algorithm includes both a Command-Line User and a Graphical User interfaces. Moreover, phonemes, classes and rules are listed in an external configuration file of the tool which can be easily modified by any user. Syllabification performance is consistent with manual annotation. This algorithm is included in SPPAS, a software for automatic speech segmentation, and distributed under the terms of the GPL license.

Italiano. *Questo articolo presenta una procedura di sillabificazione automatica per l'italiano basata su regole. Diversamente da altri sillabificatori, la nostra procedura è più facile da usare perchè l'algoritmo, compilato in Python, include un'interfaccia a linea di comando e un'interfaccia grafica. Inoltre i fonemi, le classi e le regole sono elencate in un file di configurazione esterno che può essere facilmente modificato. I risultati della sillabificazione automatica sono congruenti con quelli ottenuti dalle annotazioni a mano. L'algoritmo è incluso in SPPAS, un software per la segmentazione automatica del parlato distribuito secondo le condizioni di licenza GPL.*

1 Introduction

This paper presents an approach to automatic detection of syllable boundaries for Italian speech. This syllabifier makes use of the phonetized text.

The syllable is credited as a linguistic unit conditioning both segmental (e.g., consonant or vowel lengthening) and prosodic phonology (e.g., tune-text association, rhythmical alternations) and its automatic annotation represent a valuable tool for quantitative analyses of large speech data sets. While the phonological structure of the syllable is similar across different languages, phonological and phonotactic rules of syllabification are language-specific. Automatic approaches to syllable detection have thus to incorporate such constraints to precisely locate syllable boundaries.

The question then arises of how to obtain an acceptable syllabification for a particular language and for a specific corpus (a list of words, a written text or an oral corpus of more or less casual speech). In the state-of-the-art, the syllabification can be made directly from a text file as in (Cioni, 1997), or directly from the speech signal as in (Petrillo and Cutugno, 2003).

There are two broad approaches to the problem of the automatic syllabification: a rule-based approach and a data-driven approach. The rule-based method effectively embodies some theoretical position regarding the syllable, whereas the data-driven paradigm tries to infer new syllabifications from examples syllabified by human experts. In (Adsett et al., 2009), three rule-based automatic systems and two data-driven automatic systems (Syllabification by Analogy and the Look-Up Procedure) are compared to syllabify a lexicon.

Indeed, (Cioni, 1997) proposed an algorithm for the syllabification of written texts in Italian, by syllabifying words directly from a text. It is an algorithm of deterministic type and it is based upon the use of recursion and of binary tree in order to detect the boundaries of the syllables within each word. The outcome of the algorithm is the production of the so-called canonical syllabification (the stream of syllabified words).

On the other side, (Petrillo and Cutugno, 2003)

presented an algorithm for speech syllabification directly using the audio signal for both English and Italian. The algorithm is based on the detection of the most relevant energy maxima, using two different energy calculations: the former from the original signal, the latter from a low-pass filtered version. This method allows to perform the syllabification with the audio signal only, so without any lexical information.

More recently, (Iacoponi and Savy, 2011) developed a complete rule-based syllabifier for Italian (named Sylli) that works on phonemic texts. The rules are then based on phonological principles. The system is composed of two transducers (one for the input and one for the output), the syllabification algorithm and the mapping list (i.e., the vocabulary). The two transducers convert the two-dimensional linear input to a three-dimensional phonological form that is necessary for the processing in the phonological module and then sends the phonological form back into a linear string for output printing. The system achieved good performances compared to a manual syllabification: more than 0.98.5% (syllabification of spoken words). This system is distributed as a package written in C language and must be compiled; the program is an interactive test program that is used in command-line mode. After the program reads in the phone set definition and syllable structure parameters, it loops asking for the user to type in a phonetic transcription, calculating syllable boundaries for it, and then displaying them. When the user types in a null string, the cycling stops and execution ends. Finally, there are two main limitations: this tool is only dedicated to computer scientists, and it does not support time-aligned input data.

With respect to these already existing approaches and/or systems, the novel aspect of the work reported in this paper is as follows:

- to propose *a generic and easy-to-use tool* to identify syllabic segments from phonemes;
- to propose *a generic algorithm*, then a set of rules for the particular context of Italian spontaneous speech.

In this context, "generic" means that the phone set, the classes and the rules are easily changeable; and "easy-to-use" means that the system can be used by any user.

2 Method description

In the current study, we report on the adaptation of a rule-based system for automatic syllabification of phonemes' strings of the size greater than a graphic word. The system was initially developed for French (Bigi et al., 2010) and here adapted on Italian since there are currently no freely available system that can be used either by computer scientists and linguists.

The problem we deal with is the automatic syllabification of a phoneme sequences. The proposed phoneme-to-syllable segmentation system is based on 2 main principles:

1. a syllable contains a vowel, and only one;
2. a pause is a syllable boundary.

These two principles focus the problem on the task of finding a syllabic boundary between two vowels in each Inter-Pausal Unit (IPU), as described in Figure 1.

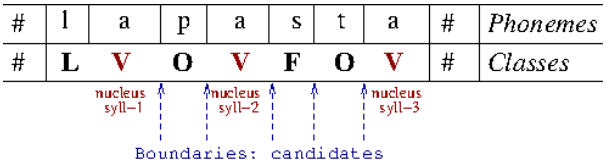

Figure 1: Main principles ("#" means a pause)

As in the initial system for French, we group phonemes into classes and establish language-specific rules dealing with these classes. The identification of relevant classes is then very important. The following classes were used:

V - Vowels: a e i o u O E oe ae
G - Glides: j w
L - Liquids: l L r
O - Occlusives: p t k b d g
F - Fricatives: s S f z tS ts v dz dZ
N - Nasals: m nf ng

Uppercase bold-letters indicate the abbreviations used for classes throughout this paper. The letter **C** is also used to mention one of G, L, O, N or F.

The system firstly check if the observed sequence of classes corresponds to an exception. If not, the general rules are applied (see Table 1).

The exception rules are:

- (Consonant + Glide) can't be segmented
- (Consonant + Liquid) can't be segmented
- (Consonant + Liquid + Glide) can't be segmented

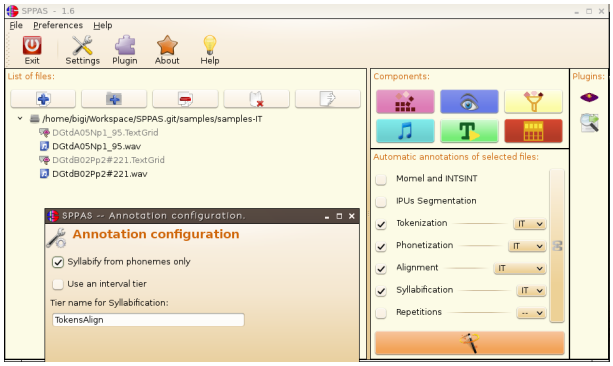

Figure 2: SPPAS: Graphical User Interface with syllabification options

	Observed sequence	Segmentation rule
1	VV	V.V
2	VCV	V.CV
3	VCCV	VC.CV
4	VCCCV	VC.CCV
5	VCCCCV	VC.CCCV
6	VCCCCCV	VCC.CCCV

Table 1: General Rules (V and C are phonological vowels/consonant respectively)

Notice that the rules we propose follow usual phonological statements for most of the spoken corpus. Our aim is not to propose a true set of syllabification rules for Italian, but to provide an acceptable syllabification for the most part of spoken corpora. We do not suggest that our solutions are the only ones particularly for syllables with complex structures as long as they are fairly uncommon in a given specific corpus. This is the reason why the tool implementing these rules was developed to be as generic as possible: any user can change either the phone set or the rules.

Finally, in the system described in (Bigi et al., 2010), the syllabification is performed between 2 silences (as defined in the main principles). From this system, we added the possibility to perform the syllabification between any kind of boundaries. In such case, a ”reference tier” is given by the user to the system. Table 2 shows an example when the time-aligned tokens are used as reference tier.

Of course, the reference tier can contain any type of annotation (we used tokens in the example, but prosodic contours, syntactic segments, etc. can be used if this annotation is available).

segment type	sentence	phonemes	syllables
sentence	la pasta	/lapasta/	la.pas.ta
	la stella	/lastela/	las.te.la
token	la.pasta	/la/ /pasta/	la.pas.ta
	la.stella	/la/ /stela/	la.ste.la

Table 2: Syllabification into segments, without changing the rules.

3 Implementing in a tool

The system proposed in this paper is included in SPPAS (Bigi, 2012), a tool distributed under the terms of the GNU Public License[1]. It is implemented using the programming language Python 2.7. Among other functions, SPPAS proposes an automatic speech segmentation at the phone and token levels for French, English, Spanish, Italian, Chinese, Taiwanese and Japanese. Moreover, the proposed software fulfills the specifications listed in (Dipper et al., 2004): it is a linguistic tool, free of charge, ready and easy to use, it runs on any platform and it is easy to install, the maintenance is guaranteed (at least until 2016), and it is XML-based. To download it, use the URL:

`http://www.lpl-aix.fr/~bigi/sppas/`

The current version (i.e. 1.6) allows to import data from Praat, Elan, Transcriber or from CSV files. The output can be one of ”xra” (native file format), ”TextGrid”, ”eaf” or ”csv”. The time-aligned phonemes (produced by SPPAS from the speech audio file and the orthographic transcription) are used as input to the syllabifier to produce 3 tiers with time-aligned syllables, classes and structures (as shown in Figure 3). A dictionary can be syllabified by using the same program, by ”simulating” time-alignments, and exporting the result in CSV format.

A simple ASCII text file that the user can change as needed contains the phoneset and the rules for the syllabification process.

4 Evaluation

All testing material was taken from CLIPS (Savy and Cutugno, 2009), distributed during the Evalita 2011 evaluation campaign. This corpus is made of about 15 map-task dialogues recorded by couples of speakers exhibiting a wide variety of Italian variants. Dialogues length ranges from 7/8 min-

[1]See: `http://www.gnu.org/licenses/gpl-3.0.en.html` for details

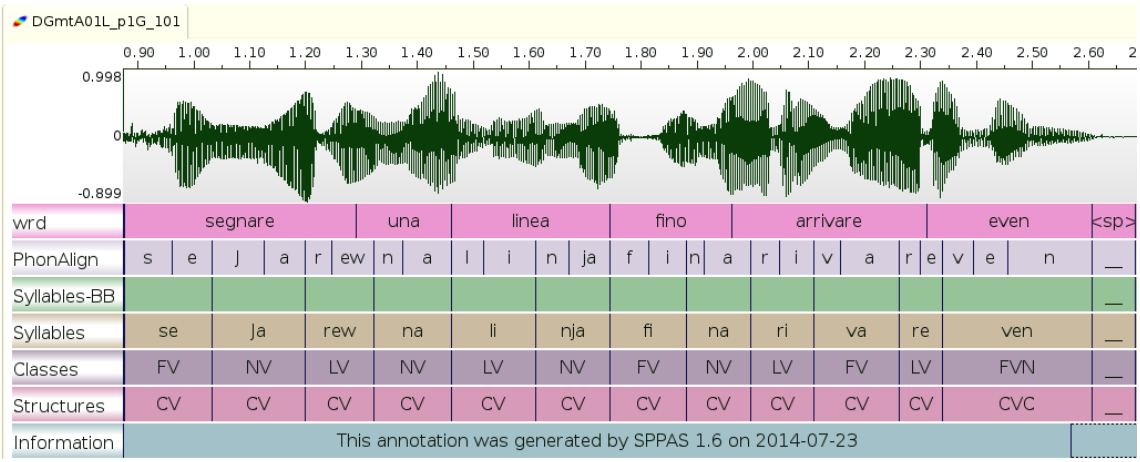

Figure 3: System output example on spoken Italian. The "Syllables-BB" tier (green) was manually annotated. Two assimilation phenomena can be observed in "fino arrivare even", which is phonetized as /finarivareven/ and that impact on the syllabification.

utes to 15/20 minutes, including word segmentation and phonetic segmentation. The test corpus was automatically extracted from these data with the following criteria: 1/ last 2 utterances of each speaker in each dialogue and 2/ all utterances containing from 100 to 106 phonemes. From such data, we kept only files containing more than 5 words, which represents about 10 minutes of spoken speech, and 1935 syllable boundaries have to be fixed. Notice that we have not corrected the transcription of phonemes for which we have not agree upon with the transcribers (as in Figure 3 for /ew/ in /seJarewna/).

The authors (one French - BB, one Italian - CP) manually syllabified the corpus and the resulting syllables were then compared with automatic syllabification obtained from the same corpus. In both cases, the syllabification was done by submitting the time-aligned phonemic representations of the sentences. One run was performed by using the basic system (phonemes only), and not by segmenting into intervals (see Figure 2 for both options). The agreement rates are:

- CP & BB: 99.12%
- CP & SPPAS-basic: 97.13%
- BB & SPPAS-basic: 97.80%

As the automatic system is using the phonemes only, it is important to notice that a part of the errors are due to the segmentation of words starting by 's' followed by a plosive (see Table 2). Unfortunately, by using the tokens as a reference tier for the segmentation, the results decrease to 96.1% (compared to BB). This is due to the large number of reductions and asimilations of spontaneous speech. However, we can create a tier with boundaries at pauses and specific boundaries before the /s/ for all words starting by /s/+plosive. The syllabification between such segments can then be used *to improve results to 98.2% (compared to CP) or 98.9% (compared to BB)*.

The results show that the program syllabification is very close to those made by human experts. Then, syllabification in Italian can be mostly predicted algorithmically, even when accounting for minor boundary segmentation phenomena found in speech.

5 Conclusion

The paper presented a new feature of the SPPAS tool that lets the user provide syllabification rules and perform automatic segmentation by means of a well-designed graphical user interface. The system is mainly dedicated to linguists that would like to design and test their own set of rules. A manual verification of the output of the program confirmed the accuracy of the proposed set of rules for syllabification of dialogues. Furthermore, the rules or the list of phonemes can be easily modified by any user. Possible uses of the program include speech corpus syllabification, dictionary syllabification, and quantitative syllable analysis.

References

Connie R Adsett, Yannick Marchand, et al. 2009. Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of italian. *Computer Speech & Language*, 23(4):444–463.

B. Bigi, C. Meunier, I. Nesterenko, and R. Bertrand. 2010. Automatic detection of syllable boundaries in spontaneous speech. In *Language Resource and Evaluation Conference*, pages 3285–3292, La Valetta (Malta).

B. Bigi. 2012. The SPPAS participation to Evalita 2011. In *Working Notes of EVALITA 2011, ISSN: 2240-5186*, Roma (Italy).

L. Cioni. 1997. An algorithm for the syllabification of written Italian. pages 22–24, Santiago de Cuba.

S. Dipper, M. Götze, and M. Stede. 2004. Simple annotation tools for complex annotation tasks: an evaluation. In *Proc. of the LREC Workshop on XML-based richly annotated corpora*, pages 54–62.

L. Iacoponi and R. Savy. 2011. Sylli: Automatic phonological syllabification for Italian. In *Proc. of INTERSPEECH*, pages 641–644, Florence (Italy).

M. Petrillo and F. Cutugno. 2003. A syllable segmentation algorithm for English and Italian. In *Proc. of INTERSPEECH*, Geneva (Switzerland).

R. Savy and F. Cutugno. 2009. CLIPS. diatopic, diamesic and diaphasic variations in spoken Italian. In *Proc. of the 5th Corpus Linguistics Conference*, Liverpool (England).