

# Annotation automatique en syllabes d'un dialogue oral spontané

Brigitte Bigi<sup>1</sup>, Christine Meunier<sup>1</sup>, Roxane Bertrand<sup>1</sup>, Irina Nesterenko<sup>2</sup>

Laboratoire Parole et Langage<sup>1</sup>,  
CNRS & Aix-Marseille Universités,  
5, avenue Pasteur,  
13100 Aix-en-Provence France,  
e-mail : brigitte.bigi@lpl-aix.fr, christine.meunier@lpl-aix.fr, roxane.bertrand@lpl-aix.fr, irina.nesterenko@inalco.fr

Institut National des Langues<sup>2</sup>  
et Civilisations Orientales  
2 rue de Lille  
75343 Paris cedex 07 France

## ABSTRACT

This paper proposes a solution to identify automatically syllable boundaries in the particular context of spontaneous speech. The main goal consists in identifying syllables from a continuous stream of phonemes. At first, phoneme classes are defined to be as well-suited as possible to reduce the problem complexity. Secondly, a few number of general rules are defined. Finally, some exception rules allows to adapt the problem to the specific context of spontaneous speech. The proposed system is evaluated and compares favorably to the only two existing other systems, for French, with significant improvements.

**Keywords :** syllable, phoneme, segmentation, rules.

## 1. INTRODUCTION

La syllabe est une des unités fondamentales de la parole et une unité structurale de grande importance dans la production et la perception du langage. Toutefois, la caractérisation de la syllabe tant sur le plan articulatoire qu'acoustique reste très difficile. Elle est néanmoins souvent considérée comme l'unité de production dans laquelle les phénomènes de coarticulation sont plus saillants [7]. Par ailleurs, elle est une unité fondamentale dans l'organisation prosodique du français, langue « syllable timed » par opposition aux langues « stressed timed » comme l'anglais [9]. Même si cette classification est parfois remise en cause [2], la réalité cognitive de la syllabe (intégrée dans la compétence des locuteurs) nécessite de la prendre en compte dans les analyses portant sur la parole continue.

Cet article s'inscrit dans le cadre du développement d'annotations multimodales de dialogues oraux spontanés. A partir de l'annotation en phonèmes, notre but consiste à proposer un outil pour identifier automatiquement les segments syllabiques. L'intérêt de ce niveau d'annotation sur *un corpus de parole continue non contrôlée* est double. Dans un premier temps, il permet d'obtenir des statistiques sur la structure et la fréquence des syllabes du français en parole non contrôlée et après la réorganisation syllabique (re-syllabation) qu'entraîne la parole continue (non dépendante du lexique) et les phénomènes de réduction spécifiques à ce type de parole. Notamment, les phénomènes de réduction tels que déformation, assimilation, élision de phonèmes, extrêmement importants en parole naturelle non contrôlée, font apparaître des séquences de segments illégaux d'un point de vue phonotactique (ex. : clusters de consonnes), mais également non prévisibles dans les combinaisons de jonction de mots. Si ces structures sont peu fréquentes, elles nécessitent néanmoins qu'on y porte intérêt. Disposer de cette annotation permet enfin d'envisager les analyses segmentales (structures acoustiques des phonèmes) et supra-segmentales (organisation rythmique et accentuelle du discours) en fonction du découpage syllabique. Par exemple, cette annotation permettra de mettre en perspective le rôle et la hiérarchie des frontières lexicales et/ou syllabiques dans la coarticulation. Ces phénomènes ont pu être évalués en parole contrôlée mais sont encore ignorés en parole non contrôlée.

Cet article concerne la proposition d'un système de détection automatique des frontières syllabiques en parole spontanée. Dans [4], quelques outils et la description du corpus ont été présentés. Le corpus - Corpus of Interactional Data (CID)<sup>1</sup>, est un enregistrement audio-vidéo de dialogues spontanés entre deux locuteurs français natifs (8 heures, 8 paires de locuteurs). Une transcription orthographique enrichie (TOE) a été réalisée et corrigée manuellement. A partir de cette TOE, un convertisseur graphème-phonème suivi d'un aligneur permettent d'obtenir une phonétisation de bonne qualité. Cette segmentation en phonèmes, rend possible la réalisation d'une recherche automatique des frontières de syllabes.

## 2. TRAVAUX EXISTANTS

A notre connaissance, il existe deux outils *libres* de recherche des syllabes à partir de phonèmes du français. *syllabation.awk*, développé par C. Pallier [8] est un script, librement diffusé, écrit en langage *awk*. Les phonèmes y sont regroupés en quatre classes : voyelles, semi-voyelles, liquides, et consonnes. Une douzaine de règles de segmentation sont définies selon les suites de phonèmes et leur classe. Ce segmenteur a été appliqué avec succès sur des mots isolés, plus spécifiquement sur les lexiques Brulex et Lexique<sup>2</sup>. Le second système libre repose également sur un ensemble de règles de segmentation appliquées sur les suites de phonèmes. *syllabify2.praat* est une partie du logiciel *EasyAlign*, développé par J-P. Goldman [6], sous la forme de scripts Praat [5]. Cinq classes de phonèmes y sont définies : voyelles, semi-voyelles, liquides, une classe comportant les consonnes p t k b d g f v et une classe comportant les consonnes s j z ʒ m n ŋ j. Ce système inclut également le silence afin de traiter la syllabation de corpus oraux. Une soixantaine de règles y ont été développées. Enfin, [1] propose une étude de la syllabation du français parlé (radio). La syllabation est réalisée avec 13 règles par l'outil GRAPHON+. Les phonèmes sont groupés en 5 classes : voyelles, semi-voyelles, liquides, occlusives et autres consonnes.

<sup>1</sup><http://crdo.up.univ-aix.fr>

<sup>2</sup><http://www.lexique.org>

### 3. SYSTÈME PROPOSÉ

#### 3.1. Principes généraux

Le système que nous avons développé s'inspire de ceux décrits précédemment en ce sens qu'il consiste à définir un ensemble de règles de segmentation entre phonèmes, selon leur classe. Cependant, avant la définition de règles, nous proposons les deux principes suivants :

**Principe 1 :** une syllabe contient une seule voyelle.

**Principe 2 :** une pause est une frontière de syllabe. Dans le corpus que nous utilisons, les pauses silencieuses supérieures à 200 ms ont été annotées automatiquement et les pauses inférieures à ce seuil ont été annotées manuellement lors de l'étape de transcription.

Ces deux principes résument le problème de syllabation que nous considérons en la recherche de frontières de syllabes entre deux voyelles.

#### 3.2. Classes de phonèmes

Les classes que nous proposons sont les suivantes :

**V** - Voyelles : i e ε a o u y ø œ ə ē ā õ

**G** - Semi-voyelles : j ɥ w

**L** - Liquides : l ʀ

**O** - Occlusives : p t k b d g

**F** - Fricatives : s z ʃ ʒ f v

**N** - Nasales : m n ŋ ɲ

La lettre en gras est le symbole utilisé dans ce papier pour désigner la classe. De plus, le symbole **X** fait mention de l'un des G, L, O, N ou F (en d'autres termes, X renvoie à un phonème qui n'est pas une voyelle).

La répartition des consonnes en trois classes est un choix majeur qui réduit largement la complexité du développement des règles. La division en un nombre plus petit de classes de consonnes impliquerait le développement, comme dans le segmenteur syllabation2.praat, d'un très grand nombre de règles pour traiter les cas particuliers. Une subdivision plus importante nous est apparue inutile car elle montrait de fortes redondances dans l'élaboration des règles, sans gain de performance.

#### 3.3. Règles de syllabation

Afin d'augmenter la généralité de l'approche, nous proposons d'organiser les règles en deux types : des règles générales décrites dans la table 1, applicables à tout type de situation, et des règles exceptions qui peuvent être modulées selon le problème abordé (dialogue spontané ou non, etc.) décrites dans la table 2.

La première règle générale applique le principe 1 selon lequel il n'y a qu'une voyelle par syllabe. La deuxième règle reflète la tendance universelle à privilégier les syllabes ouvertes, en conséquence de quoi la consonne est assignée à la seconde syllabe. Les règles générales 4, 5 et 6 satisfont la règle du « Maximum Onset Principle » pour laquelle, dans un groupe de consonnes intervocaliques, le maximum de consonnes doivent être attribuées à l'onset de la seconde syllabe plutôt qu'à la coda de la première. La troisième règle doit être considérée en fonction des règles exceptions 1, 2 et 3 (table 2), dans lesquelles les clusters intervocaliques sont constitués de deux consonnes. Ces

TAB. 1: Règles générales

	Séquence	Règle	Exemples
1	VV	V.V	poète : po.ɛt il y a un : i.a.œ̃ en haut : ā.o
2	VXV	V.XV	limité : li.mi.te et donc on : e.dõ.kõ
3	VXXV	VX.XV	jardin : ʒa.r.dẽ comme ça : kom.sa parce qu'il : pas.ki
4	VXXXV	VX.XXV	avec moi : a.vek.mwa cheval noir : sɔ.val.nwa.r
5	VXXXXV	VX.XXXV	il se présentait : il.spre.zã.te
6	VXXXXXV	VXX.XXXV	alors je crois : a.lorʒ.krwa

TAB. 2: Règles exceptions

	Séquence	Règle	Exemples
1	VXGV	V.XGV	baaignoire : be.nwa.r spéciaux : spe.sjo tu vois : tu.vwa
2	VFLV	V.FLV	découvre : de.ku.vrɔ̃,
3	VOLV	V.OLV	il trouve : i.truv mais de la : me.dla
4	VFLGV	V.FLGV	effroyable : e.frwa.jabl
5	VOLGV	V.OLGV	incroyable : ê.krwa.jabl
6	VOLOV	VOL.OV	connaître tu : ko.netr.ty capable parce : ka.pabl.pas

règles montrent que pour un cluster de deux consonnes le « Maximum Onset Principle » doit être appliqué prudemment et en fonction du principe de sonorité, c'est-à-dire de la nature des consonnes du cluster. Ainsi, la règle générale ne s'applique que lorsque le principe de sonorité est violé au sein du cluster.

Les règles exceptions 4 et 5 sont liées au statut particulier des clusters Obstruante + Liquide + Glide en français, ces groupes étant le plus souvent homosyllabiques. La règle exception 6 est une exception au « Maximum Onset Principle ». Cette exception est largement motivée par la nature de la parole continue et le fait qu'il n'existe pas, à notre connaissance, de modèles permettant de combiner systématiquement frontière syllabique et frontière lexicale. Ainsi, un cluster tel que Plosive + Liquide + Plosive n'existe pas en interne de mot en français ; malgré tout il apparaît dans notre corpus en raison des phénomènes de réduction caractéristiques de ce type de parole.

Les règles que nous proposons suivent les principes usuels bien connus dans le domaine de la phonologie et peuvent être appliqués à notre corpus comme à d'autres. En ce sens, notre but n'est pas de proposer des règles spécifiques à notre corpus mais bien un principe général, possiblement adaptable à tout autre type de corpus.

Finalement, ce travail a été implémenté sous forme de classes java. Le programme utilise un fichier de configuration qui décrit la liste des phonèmes et leur classes, ainsi que la liste de toutes les règles. Ces paramètres peuvent ainsi être facilement modifiés. La version diffusée en GPL,

nommée *LPL-Syllabeur-v2.1.jar* prend en entrée un fichier *praat* de phonèmes encodés en SAMPA et rend en sortie un fichier *praat* de syllabes.

## 4. EVALUATION

### 4.1. Description du corpus

Le système décrit dans ce papier a été utilisé pour l'annotation du Corpus of Interactional Data (CID) [3, 4]. Le CID est un corpus de 8 heures d'enregistrements audio et vidéo de dialogues français en parole spontanée (1 heure d'enregistrement par session). Chaque dialogue est constitué de deux participants du même sexe qui étaient amenés à converser autour de l'un des deux thèmes suivants : les conflits dans leur environnement professionnel ou bien les situations insolites qu'ils ont pu connaître. Ces consignes n'étaient pas strictes et les participants ont souvent fait évoluer la conversation vers d'autres thématiques.

L'une des caractéristiques majeure de la parole spontanée est l'écart manifeste que l'on peut observer entre des réalisations standards (« orthographic token ») et la production réelle des locuteurs. Les élisions et réductions telles que « je suis » produit [ʃi] ou « je ne sais pas » produit [ʃepa] sont extrêmement fréquentes et peuvent être extraites d'un lexique de variantes prototypiques. Mais un corpus conversationnel tel que le CID présente de surcroît de nombreuses variantes non prototypiques qu'il est illusoire de vouloir stocker dans un lexique ([3]). Ces phénomènes rendent la détection automatique de syllabe tout à fait spécifique. Par conséquent, et compte-tenu de l'aspect modulaire du syllabeur proposé, nous avons décidé de résoudre les cas particuliers les plus fréquents de la manière suivante : les enchaînements de phonèmes suivants fs, pt, sk (excepté quand pVsk) n'ont pas été éclatés, car ils correspondent à des unités lexicales très fréquentes (pe-tit, parce que, puisque, faisait, etc).

### 4.2. Protocole d'évaluation

Dans cette section, nous présentons les résultats d'une comparaison établie entre deux syllabations manuelles et la syllabation automatique du CID. Le corpus de test représente environ 7 minutes d'un dialogue, soit environ 2000 mots (653 d'un locuteur, 1238 du second). Pour les évaluations, nous avons choisi de ne pas prendre en considération les deux principes énoncés dans la section 3.1. Selon le premier principe, les séquences VV ont une segmentation évidente qu'il n'est pas utile d'évaluer. De même, dans l'estimation des performances, nous n'avons pas inclus les cas concernés par le second principe : une voyelle suivie d'une pause, une pause suivie par une autre pause et une pause suivie d'une voyelle.

Les évaluations portent sur 1646 frontières pour lesquelles une décision doit être prise. La table 3 apporte des précisions sur les règles qui sont utilisées et pour lesquelles le système est évalué.

Les segmentations manuelles ont été réalisées par deux annotateurs qui disposaient des phonèmes, et de la transcription. Il est important de rappeler ici que contrairement aux annotateurs, le système automatique ne s'appuie que sur les phonèmes, sans la transcription orthographique. Le taux d'accord inter-annotateur est de 98,60 %, ce qui signifie que 23 frontières proposées sont différentes, sur les

**TAB. 3:** Statistiques sur l'utilisation des règles (1646 frontières)

Nombre	Règle		Nombre	Règle
1165	VXV			
435	VXXV	dont	54	VXGV
			17	VFLV
			73	VOLV
43	VXXXV	dont	0	VFLGV
			4	VOLGV
			4	VOLOV
3	VXXXXV			

1646 possibles. Ces 23 désaccords se répartissent différemment selon le nombre de consonnes qui sépare deux voyelles : plus leur nombre est important, plus le désaccord augmente, comme on le constate ci-après :

- 5 désaccords en VXV, soit 0,43 % des 1165 cas,
- 12 désaccords en VXXV, soit 2,76 % des 435 cas,
- 5 désaccords en VXXXV, soit 11,63 % des 43 cas,
- 1 désaccord en VXXXXV, soit 33,33 % des 3 cas.

### 4.3. Qualité de la syllabation

Afin de situer notre proposition par rapport aux systèmes existants, nous avons (1) implémenté les règles de C. Pallier dans un fichier de configuration de notre système, (2) idem pour les règles de la table 1 de [1] et (3) adapté l'outil de J-P. Goldman de sorte qu'il soit applicable à notre corpus (en particulier l'encodage des phonèmes). Enfin, nous avons évalué l'ensemble des outils sur le corpus de test.

La table 4 montre l'ensemble des performances des systèmes automatiques, par rapport à chacun des annotateurs (mentionnés Annot1 et Annot2). On y constate que la syllabation que nous proposons offre un gain relatif d'environ 30-35 % par rapport aux systèmes de l'état de l'art. Plus encore que ce gain de performance, l'outil que nous avons développé a vocation à être modulable à volonté en fonction du contexte (encodage des phonèmes, formats des fichiers, etc.), et diffusé librement.

**TAB. 4:** Nombres de désaccords et pourcentages

	Annot. 1	Annot. 2
syllabation.awk (de [8])	74 4,50 %	84 5,10 %
graphon+ (de [1])	85 5,16 %	92 5,59 %
syllabify2.praat (de [6])	67 4,07 %	75 4,56 %
LPL-syllabeur-v2.1.jar (système proposé)	43 2,61 %	53 3,22 %

### 4.4. Analyse des différences

Comme présenté en table 3, on observe que 97,21 % des frontières syllabiques sont concernées par les règles suivantes : VXV, VXXV, VXGV, VOLV et VFLV. Ces règles ne présentent pas d'ambiguïté (voir les détails en table

5). En conséquence, les résultats de la syllabation automatique devraient être totalement conformes aussi bien à une segmentation selon des règles phonotactiques qu'aux intuitions des auditeurs. Le problème majeur pour la syllabation reste les cas où deux voyelles sont séparées par plus de deux consonnes. Ces occurrences sont rares (2,61 %) et la plus fréquente est VXXXV (dans laquelle sont exclus les cas où C2 est une plosive ou une fricative). Notre proposition (à l'instar du système de règles de Goldman) est de mettre la frontière syllabique entre C1 et C2. Les éventuelles erreurs de syllabation seront corrigées après le passage d'un expert.

**TAB. 5:** Désaccords entre le système proposé et les annotateurs

	Annot. 1	Annot. 2
VXV	5 0,43 %	4 0,34 %
VXXV+exceptions	26 5,98 %	32 7,36 %
VXXXV+exceptions	11 25,59 %	15 34,88 %
VXXXXV	1 33,33 %	2 66,67 %
Total	43	53

Nous proposons ci-après (table 6) quelques exemples des syllabations produites par notre système comparativement à celles des annotateurs. Il est à noter qu'une grande partie des différences observées entre les syllabations manuelle et automatique sont concentrées aux jonctures de mots. Il semble que les annotateurs humains soient influencés par les frontières lexicales lorsque la syllabation est complexe, comme « parcs c'est » dans l'exemple. Le syllabeur ne tient pas compte des frontières lexicales car il ne dispose pas de cette information. Il n'est pas question de décider *a priori* quelle est la meilleure syllabation, toutefois, l'information fournie par le syllabeur (sans influence lexicale) offre la possibilité d'évaluer les rôles des frontières lexicales et syllabiques indépendamment dans nos futures analyses phonétiques.

## 5. CONCLUSION

Dans cet article, nous avons proposé un outil permettant la syllabation, à partir de phonèmes, d'un corpus de dialogue oral spontané. Ses performances ont été évaluées et jugées satisfaisantes pour les tâches ultérieures auxquelles il est dédié. Nous avons regroupé les phonèmes dans des classes et défini un ensemble de règles générales suivies d'exceptions afin de trouver les frontières pertinentes entre les syllabes. Cet outil est destiné à être appliqué sur l'ensemble du CID afin d'être mis en relation avec les autres niveaux d'annotation existants. Mais ce niveau d'annotation va permettre également d'effectuer de nouvelles annotations telles que celles par exemple de la structure syllabique (onset, noyau, coda) qui n'a pas été beaucoup étudiée sur ce type de parole.

## RÉFÉRENCES

[1] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their

**TAB. 6:** Exemple de syllabation manuelle et automatique

Transcription	et donc on mange sur la baignoire donc c'est c'est ça
Phonèmes	e d ð k ð m ã ʒ s y r l a b e n w a r d ð k s e s e s a
Classes	V O V O V N V F F V L L V O V N G V L O V O F V F V F V
Syllabes (Auto & Annot)	e . d ð . k ð . m á ʒ . s y r . l a . b e . n w a r . d ð k . s e . s e . s a
Transcription	non dans les parcs c'est un peu limité
Phonèmes	n ð d ð l e p a r k s e t æ p ə l i m i t e
Classes	N V O V L V O V L O F V O V O V L V N V O V
Syllabes Auto	n ð . d ð . l e . p a r . k s e . t æ . p ə . l i . m i . t e
Syllabes Annot1 & Annot2	n ð . d ð . l e . p a r k . s e . t æ . p ə . l i . m i . t e
Transcription	il expliquait pas vraiment ce qu'il y avait dedans
Phonèmes	i l e k s p l i k e p a v r e m ä s k i j a v e d ä
Classes	V G V O F O L V O V O V F L N V N V F O V G V F V O V
Syllabes Auto	i . l e k . s p l i . k e . p a . v r e . m ä . s k i . j a . v e . d ä
Syllabes Annot1	i . l e k . s p l i . k e . p a . v r e . m ä . s k i . j a . v e . d ä
Syllabes Annot2	i . l e k s . p l i . k e . p a . v r e . m ä . s k i . j a . v e . d ä

variation in spontaneous french. *Speech Communication*, 46 :119–139, 2005.

- [2] C. Astesano. *Rythme et accentuation en Français*. L'Harmattan, coll. Langue et parole, Invariance et variabilité stylistique, 2001.
- [3] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. Le cid - corpus of interactional data. *Traitement Automatique des Langues*, 49(3) :105–134, 2008.
- [4] P. Blache, R. Bertrand, and G. Ferré. Creating and exploiting multimodal annotated corpora : the toma project. *Multimodal Corpora*, LNAI 5509 :38–53, 2009.
- [5] P. Boersma and D. Weenink. Praat : doing phonetics by computer, <http://www.praat.org>.
- [6] J-P. Goldman. <http://latlcui.unige.ch/phonetique/>, 2007.
- [7] V-A. Kozhevnikov and L-A. Chistovitch. *Speech Articulation and Perception*, volume 30. Washington D.C. : Joint Publications Research Service, 1965.
- [8] C. Pallier. Syllabation des représentations phonétiques de brulex et de lexique, <http://www.pallier.org>, 1999.
- [9] K-L. Pike. *The intonation of American English*. The University of Michigan Press in Linguistics 1, 1945.