SyllabO : une nouvelle base de données des syllabes du français québécois oral et leurs statistiques distributionnelles

LABORATOIRE DES NEUROSCIENCES DE LA PAROLE ET DE L'AUDITION
SPEECH AND HEARING NEUROSCIENCE LABORATORY

Pascale Bédard[1,2], Patrick Drouin[3], Johanna-Pascale Roy[4], Anne-Marie Audet[1,2], Julie Rivard[1,2], Pascale Tremblay[1,2]

1: Université Laval, Département de réadaptation, Québec, QC. 2: Centre de recherche de l'Institut universitaire en santé mentale de Québec, Québec, QC.
3: Université de Montréal, Département de linguistique et de traduction, Montréal, QC. 4: Université Laval, Département de langues, linguistique et traduction, Québec, QC.

Institut universitaire en santé mentale de Québec
UNIVERSITÉ LAVAL

Introduction

La communication est au cœur de notre vie et repose en grande partie sur notre habileté à parler. Le langage parlé est très complexe et à l'intérieur de celui-ci se trouvent des quantités de statistiques distributionnelles. En effet, la fréquence d'occurrence et les mesures d'association entre les sons, les syllabes et les mots qui constituent la langue varient énormément. Toutes ces statistiques (p. ex.: fréquence) sont spécifiques selon chaque langue (p. ex.: anglais « th », français « é »), selon chaque modalité (orale, écrite) et selon chaque contexte de communication (formel, informel).

La syllabe joue un rôle primordial car elle est l'unité de base en perception et en production de la parole (1, 2, 3). Elle est importante au cours du développement langagier (4, 5) et des études ont montré des réseaux cérébraux spécifiques au traitement de la syllabe (6, 7, 8, 9, 10, 11). La fréquence des syllabes a un effet facilitateur sur la production du langage; les syllabes plus fréquentes sont produites plus rapidement et avec moins d'erreurs (12, 13, 14). De plus, les adultes comme les enfants utilisent les probabilités de transition des syllabes pour apprendre à segmenter les sons dans le flot continu de la parole (15, 16, 17, 18, 19, 20).

Problématique

Afin de comprendre les mécanismes de perception et de production de la parole, il est nécessaire de comprendre l'utilisation naturelle de la syllabe, dans la vie de tous les jours. Toutefois, aucune base de données sur les statistiques distributionnelles des syllabes dans le français québécois oral n'était disponible à ce jour.

Objectifs

Le but du projet était donc de créer un nouvel outil pour l'étude du langage parlé, d'abord par la création d'un corpus du français québécois oral (objectif 300 000 syllabes), puis d'une base de données et d'outils d'analyse.

Résultats & Discussion

Projet complété avec succès – plus de 300 000 syllabes composent la base de données SyllabO

Syllabes

Tableau 1. Nombre de syllabes et nombre de locuteurs (N) en fonction du contexte, de l'âge et du genre

Âge / Genre / Contexte	18-45 ans Homme	18-45 ans Femme	46-70 ans Homme	46-70 ans Femme	71-100 ans Homme	71-100 ans Femme	Total
Formel	27126 N = 12	28744 N = 12	27324 N = 11	27381 N = 10	25043 N = 11	25978 N = 12	161596 N = 68
Informel	25527 N = 25	25682 N = 22	25100 N = 22	25941 N = 21	25659 N = 14	14131 N = 12	142040 N = 116
Total	52653 N = 37	54426 N = 34	52424 N = 33	53322 N = 31	50702 N = 25	40109 N = 24	**303636** N = 184

Le moteur de recherche permet d'obtenir les tableaux résultats à partir du corpus complet, où à partir d'un sous-ensemble spécifique en filtrant les paramètres de contexte, d'âge et de genre.

Tableaux de la base de données

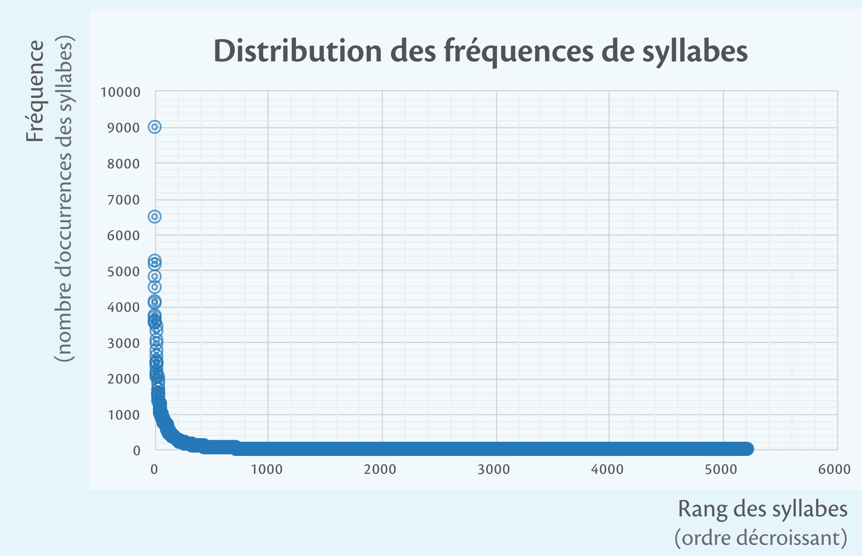
Syllabe	Structure	Fréquence	Pourcentage	Rang centile
a	V	8994	2.96209935581	99.9808171878
se	CV	6497	2.13973310148	99.9616343756
de	CV	5245	1.72739727832	99.9424515634
də	CV	5156	1.69808586597	99.9232687512
le	CV	4802	1.5814989	99.904085939
e	V	4510	1.48533112016	99.8849031268
lɑ	CV	4133	1.36116929481	99.8657203146
la	CV	4107	1.35260641031	99.8465375024
mɑ̃	CV	3763	1.23931286145	99.8273546902

Les données sont présentées sous forme de tableaux, avec les valeurs associées aux syllabes ou aux groupes de syllabes (paires, triades, tétrades). On y retrouve la structure syllabique (consonnes / voyelles), la fréquence, le rang centile et les mesures d'association (probabilités de transition, PMI et MI-like).

Paire de syllabes	Fréquence	Pourcentage	Rang centile	Probabilité de transition avant	Probabilité de transition arrière	Score d'association (PMI)
a vɛk	932	0.312041730559	99.9987523238	10.3624638648	95.6878850103	5.01364433951
a lɔr	662	0.22164337514	99.9975046476	7.36046253058	91.8169209431	4.9540681145
se sɑ	567	0.189836546381	99.9950092952	8.72710481761	19.2529711375	3.16957834198
pars kə	567	0.189836546381	99.9950092952	67.4197384067	15.3950583763	5.79656057839
vrɛ mɑ̃	545	0.182470754458	99.993761619	74.7599451303	14.4831251661	5.91465316688
bu ku	524	0.175439771259	99.9925139428	74.8571428571	51.9326065411	7.81548475621
u si	496	0.166065126993	99.9912662666	32.9349269588	16.1721552005	5.02706113005
e te	490	0.16405627465	99.9900185904	10.8647450111	13.88101983	3.22425707831
ɔ̃ nɑ	448	0.149994308252	99.987523238	13.6377473364	53.8461538462	5.63722087431

Statistiques distributionnelles

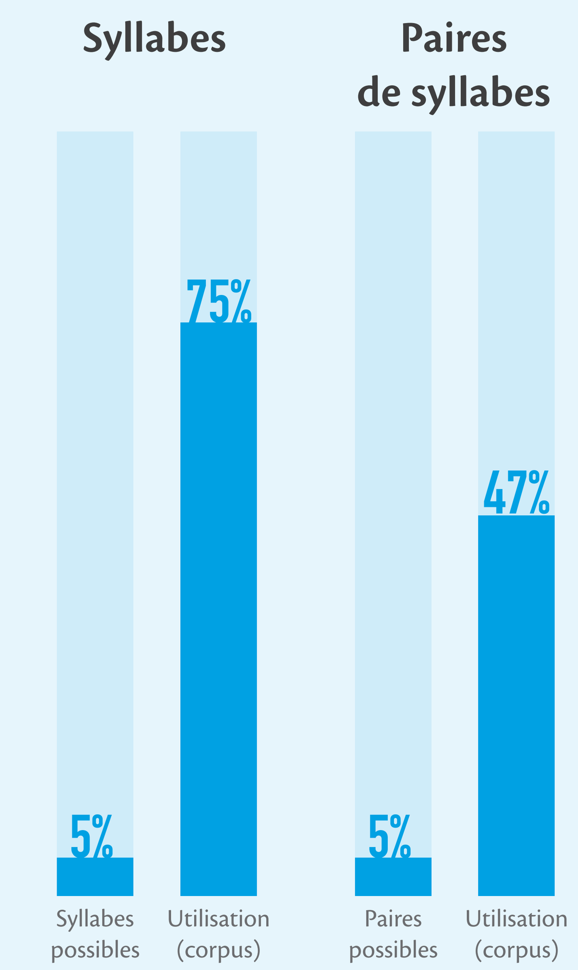
Quelques résultats saillants illustrés à partir des informations de la base de données


Distribution des fréquences de syllabes
Fréquence (nombre d'occurrences des syllabes)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
Rang des syllabes (ordre décroissant)

Il y a peu de syllabes très fréquentes et un grand nombre de syllabes rares.

Structures syllabiques: consonnes – voyelles

Il y a une prévalence de structures syllabiques complexes (CCV, CCVC, CVC) parmi toutes les syllabes possibles, mais les syllabes les plus utilisées dans le corpus sont celles qui ont une structure simple CV (52% de l'utilisation).


Syllabes
Paires de syllabes
5%
75%
5%
47%
Syllabes possibles
Utilisation (corpus)
Paires possibles
Utilisation (corpus)

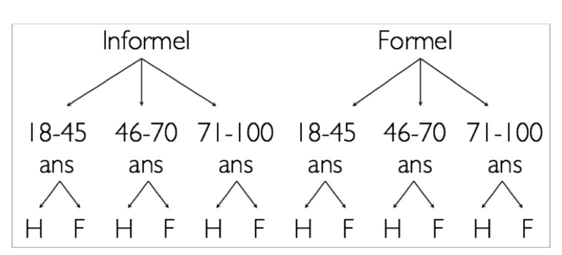
Un petit nombre de syllabes et de paires de syllabes sont utilisées très fréquemment dans le corpus.

Méthodes


1
2
[dø]
3
4
<xml>
5
.py
6
7

1. Enregistrements audio
- Locuteurs québécois (184)
- Hommes (95) / femmes (89)
- Discours spontané
- Contextes formel / informel
- Thèmes variés (systématique)

Répartition des enregistrements (contexte, âge, genre) pour assurer la représentativité


Informel
Formel
18-45 ans
46-70 ans
71-100 ans
18-45 ans
46-70 ans
71-100 ans
H F H F H F H F H F H F

2. Transcription
- Protocole de transcription
- Alphabet phonétique international (API)
 Un son = un symbole
 Exemples : « é », « ai », « er » – [e]
 « c » – [s] ou [k]
- Règles précises respectant les phénomènes du français québécois oral
- Uniformité

- Exemple
 De l'écoute de:
 « Elle étudie à l'université… »
 À la transcription:
 [ɛl etydi a lynivɛrsite]

3. Syllabation
- Protocole de syllabation
- Découpage de la transcription en unités individuelles (syllabes)
- Règles respectant les phénomènes de l'oral tels que les liaisons, les épenthèses et l'élision
- Accord interjuge:
 91% et plus pour le discours informel
 94% et plus pour le discours formel

4. Données
- Organisation de ces transcriptions syllabées

5. Fichiers XML
- Facilite la structuration et l'utilisation des données
- XML : type de fichiers avec balises (arborescentes)
- Permet le transport efficace d'une grande quantité de données
- Un fichier par transcription
- Un fichier de métadonnées (anonymisation)

6. Script
- Création d'un script Python
- Algorithme pour chaque information à extraire
- Organisation de syllabes et de leurs statistiques sous forme de base de données

7. Base de données
- Moteur de recherche
 - Sélection selon les métadonnées (âge, genre, contexte)
 - Génération de tableaux pour chaque groupe de 1 à 4 syllabes
 1 syllabe - [la]
 2 syllabes - [ki] [pø]
 3 syllabes - [a] [ly] [ni]
 4 syllabes - [fi] [kyl] [te] [a]
 - Données:
 - Structure syllabique (consonnes / voyelles)
 - Fréquence
 - Fréquence en pourcentage
 - Rang centile
 - Probabilités de transition
 - Information mutuelle (PMI et MI-like)
- Application web (équipe de développeurs)

Conclusion

La création du corpus et de SyllabO (base de données et outil de recherche) a été complétée avec succès. Au terme du projet, SyllabO sera accessible gratuitement pour téléchargement à partir du site web de notre Laboratoire (www.speechneurolab.ca).

Nous espérons que SyllabO soit utile dans plusieurs champs d'application:
- **Neurosciences cognitives du langage et psycholinguistique**
 Création de stimuli mieux contrôlés pour l'étude de la production et de la perception de la parole aux niveaux comportemental et neurologique
- **Orthophonie**
 Définir des cibles d'intervention plus précises et potentiellement plus efficaces en se basant sur les statistiques d'utilisation des syllabes dans le discours parlé au Québec
- **Phonétique et phonologie**
 Description de l'utilisation des syllabes en français québécois oral, caractérisation en terme de complexité syllabique et étude des règles de syllabation

Des perspectives encore plus large sont envisagées pour SyllabO, avec l'ajout de bases de données lexicale et phonémique.

Réferences

[1] Levelt et al., 1999. [2] Indefrey & Levelt, 2004. [3] Kent, 2000. [4] Davis & Zajdo, 2010. [5] Hoff & Shatz, 2009. [6] Bohland & Guenther, 2006. [7] Peeva et al 2010. [8] Bohland et al., 2010. [9] Brendel et al., 2011.[10] Deschamps & Tremblay, 2014. [11] Tremblay & Deschamps 2015. [12] Levelt, 1999. [13] Carreiras et al., 2006. [14] Cholin et al., 2006. [15] Newport & Aslin, 2004. [16] Pelucchi, Hay, & Saffran, 2009a, 2009b. [17] Pena, Bonatti, Nespor, & Mehler, 2002. [18] Saffran, Aslin, & Newport, 1996. [19] Saffran, Johnson, Aslin, & Newport, 1999. [20] Tremblay et al., 2012.

Remerciements

CRSNG NSERC
INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada