

# *SyllabO+: A new tool to study sublexical phenomena in spoken Quebec French*

**Pascale Bédard, Anne-Marie Audet,  
Patrick Drouin, Johanna-Pascale Roy,  
Julie Rivard & Pascale Tremblay**

**Behavior Research Methods**

e-ISSN 1554-3528

Behav Res

DOI 10.3758/s13428-016-0829-7



## **Behavior Research Methods**

VOLUME 48, NUMBER 4 ■ DECEMBER 2016

# BRM

**EDITOR**

Michael N. Jones, *Indiana University*

**ASSOCIATE EDITORS**

Dale Barr, *University of Glasgow*

Amy H. Criss, *Syracuse University*

Rick Dale, *University of California, Merced*

Chris Donkin, *University of New South Wales*

Mark W. Greenlee, *University of Regensburg*

Daniel Navarro, *University of Adelaide*

Melvin J. Yap, *National University of Singapore*

A PSYCHONOMIC SOCIETY PUBLICATION

[www.psychonomic.org](http://www.psychonomic.org)

ISSN 1554-3528

 Springer



**Your article is protected by copyright and all rights are held exclusively by Psychonomic Society, Inc.. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at [link.springer.com](http://link.springer.com)".**

# SyllabO+: A new tool to study sublexical phenomena in spoken Quebec French

Pascale Bédard<sup>1,2</sup> · Anne-Marie Audet<sup>1,2</sup> · Patrick Drouin<sup>3</sup> · Johanna-Pascale Roy<sup>1,4</sup> · Julie Rivard<sup>1,2</sup> · Pascale Tremblay<sup>1,2</sup> 

© Psychonomic Society, Inc. 2016

**Abstract** Sublexical phonotactic regularities in language have a major impact on language development, as well as on speech processing and production throughout the entire lifespan. To understand the impact of phonotactic regularities on speech and language functions at the behavioral and neural levels, it is essential to have access to oral language corpora to study these complex phenomena in different languages. Yet, probably because of their complexity, oral language corpora remain less common than written language corpora. This article presents the first corpus and database of spoken Quebec French syllables and phones: *SyllabO+*. This corpus contains phonetic transcriptions of over 300,000 syllables (over 690,000 phones) extracted from recordings of 184 healthy adult native Quebec French speakers, ranging in age from 20 to 97 years. To ensure the representativeness of the corpus, these recordings were made in both formal and familiar communication contexts. Phonotactic distributional statistics (e.g., syllable and co-occurrence frequencies, percentages, percentile ranks, transition probabilities, and pointwise mutual information) were computed from the corpus. An open-access

online application to search the database was developed, and is available at [www.speechneurolab.ca/syllabo](http://www.speechneurolab.ca/syllabo). In this article, we present a brief overview of the corpus, as well as the syllable and phone databases, and we discuss their practical applications in various fields of research, including cognitive neuroscience, psycholinguistics, neurolinguistics, experimental psychology, phonetics, and phonology. Nonacademic practical applications are also discussed, including uses in speech–language pathology.

**Keywords** Syllable · Corpus · Oral language · Phonotactic regularities · Distributional statistics · Transition probabilities · Phones

In this article, we introduce a new and unique corpus of spoken Quebec French and the two sublexical databases that were derived from it to enable the study of the distributional properties of Quebec French sublexical units. Sublexical units of language, such as syllables, consonant clusters, phonemes, or phones, have distributional properties, such as co-occurrence frequency and transition probabilities, that influence language development, as well as language processing and production throughout the entire lifespan. The syllable, often considered the basic unit of speech perception and production (Levelt, 1999), plays a fundamental role in child language development. Indeed, researchers have shown that complex syllable structures (i.e., those including a consonant cluster) are acquired later than simple syllables (with a consonant vowel [CV] syllabic structure; Levelt, Schiller, & Levelt, 2000; Lleó & Prinz, 1996; McLeod, van Doorn, & Reed, 2001). Complex syllables are also associated with more articulation errors in patients with speech apraxia, a speech motor programming disorder, who tend to delete consonants to simplify syllabic structure (e.g., Aichert & Ziegler, 2004; Romani, Galluzzi, Bureca, & Olson, 2011).

**Electronic supplementary material** The online version of this article (doi:10.3758/s13428-016-0829-7) contains supplementary material, which is available to authorized users.

✉ Pascale Tremblay  
pascale.tremblay@fmed.ulaval.ca

<sup>1</sup> Département de Réadaptation, Université Laval, 1050 avenue de la Médecine, Québec, Québec G1V 0A6, Canada

<sup>2</sup> Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec (CRIUSMQ), Québec, Québec, Canada

<sup>3</sup> Département de linguistique et traduction, Université de Montréal, Montréal, Québec, Canada

<sup>4</sup> Département de Langues, Linguistique et Traduction, Université Laval, Québec, Québec, Canada

In addition to syllable structure, syllable distribution also affects language processing and production. Specifically, it has been suggested that children use information about the distribution of syllables in a language (including transition probabilities) to learn words. This is useful given that oral languages do not contain pauses marking word boundaries (unlike written languages that use blank spaces). However, syllables that frequently co-occur have high *transition probabilities* and they tend to form words. This information can be used to segment words from the speech signal. Experimental evidence shows that children are sensitive to syllable distributional statistics (Goyet, Nishibayashi, & Nazzi, 2013; Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009). Specifically, they can use transition probabilities to learn to extract words from the continuous speech flow in a new language (Pelucchi, Hay, & Saffran, 2009a, 2009b) or in an artificial language composed of nonwords (Saffran, Aslin, & Newport, 1996). Importantly, adults are also sensitive to the statistical distribution of linguistic units, including in series of syllable sequences (Newport & Aslin, 2004; Peña, Bonatti, Nespore, & Mehler, 2002), and in words and nonwords in which phonemes vary according to their phonotactic probabilities (Vitevitch, 2003; Vitevitch & Luce, 1998; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Vitevitch, Luce, Pisoni, & Auer, 1999). In addition to being sensitive to transition probabilities, a facilitating effect of syllable frequency on speech production has also been observed in healthy adults: frequent syllables are produced more rapidly and more accurately (e.g., Cholin, Levelt, & Schiller, 2006; Levelt, 1999). Similarly, it has also been shown that patients with apraxia of speech make more errors in words containing a less-frequent first syllable (e.g., Aichert & Ziegler, 2004; Staiger & Ziegler, 2008). In addition to the behavioral evidence, a growing body of neuroimaging studies has shown sensitivity to speech statistical information in a number of brain regions, including the inferior frontal gyrus, supratemporal cortex, and ventral premotor cortex, during the experimental manipulation of transition probabilities (Deschamps, Hasson, & Tremblay, 2016; Karuza, Newport, Aslin, Starling, Tivarus, & Bavelier, 2013; Leonard, Bouchard, Tang, & Chang, 2015; Tremblay, Baroni, & Hasson, 2012; Tremblay, Deschamps, Baroni, & Hasson, 2016), phonotactic frequencies of phonemes (i.e., co-occurrence frequency; Vaden et al., 2011; Vaden, Piquado, & Hickok, 2011), or mutual information (Tremblay et al., 2016).

In brief, even though questions remain regarding the role of sublexical distributional statistics in language processing and production, it is clear that they influence both behavior and brain activity throughout life, and because of this, it is critical that researchers have access to tools that will allow them to study these effects in as many languages as possible. This is important because the concept of distributional properties is universal. That is, all languages are composed of sublexical units that vary in their frequencies of use, and that are assembled to form words and sentences following

“rules” or regularities that determine the permissible combinations of phonemes, phones, and syllables. However, because each language is composed of a different set of sublexical units that are organized according to a number of language-specific phonotactic and syntactic rules, the actual distributional properties associated with any given sublexical unit are language-specific, even though the same syllables may actually occur in several languages. For example, the syllable [das] is present in both German and Italian, but its distributional statistics are different, notably in terms of frequency (in percentages): whereas the German syllable frequency<sup>1</sup> is 1.6247 %, the Italian syllable frequency<sup>2</sup> is much lower, only 0.0023 %.

Moreover, even within a language, geographical variations can often differ significantly, in terms of both phonetic inventory and vocabulary use, thereby affecting sublexical distributional statistics. This is the case for French, the language of interest in this article. French from Quebec and France (and other geographical varieties) differ in terms of word use and phonological inventory (e.g., Ciolac, 2010; Gess, Lyche, & Meisenburg, 2012; Klein & Rossari, 2003). For example, the phoneme [œ] (e.g., *quelqu'un*, “someone” [kɛlkœ] or *brun*, “brown” [brœ]) is frequently used by French speakers in Quebec, and yet it is now hardly ever used by French speakers from France, who favor [ɛ̃] (e.g., *kɛlkɛ̃* or [brɛ̃]) (e.g., Akamatsu, 1967; Canepari, 2005; Martin, Beaudoin-Begin, Goulet, & Roy, 2001; Vajta, 2012). Thus, the distributional statistics of both phonemes ([œ] and [ɛ̃]), as well as the syllables that include these phonemes, are likely to differ widely across the two varieties of French. Another example is the word *char* (“car”). This word is used by speakers of French in both Quebec and France (and most likely other varieties of French, too). However, in France, the use of *char* is limited to the rather infrequent expression *char d'assaut* (“tank”), whereas in Quebec it is commonly used, in informal oral contexts, to refer to nonmilitary vehicles (cars), a far more familiar notion (Mercier, 2002). Hence, the frequency of the syllable [ʃar] (*char*) probably differs widely across the two regions. Thus, geographic—diatopic—(Moreau, 1997) varieties of French, such as Quebec French, France French, Belgium French, and New-Brunswick French, can be distinguished in terms of sounds, syllables, or word inventories, as well as in terms of the distribution of each unit. Hence, even if a large number of units are shared across geographical varieties of a language, the distributional statistics are unique to each variety, and research tools aiming to provide information about frequency of use and other distributional information should take into account the variety of interest (Poděšva & Sharma, 2014).

<sup>1</sup> Retrieved from the BASTat database (Schiel, 2010).

<sup>2</sup> Retrieved from the PhonItalia database (Goslin, Galluzzi, & Romani, 2013).

Another important factor when studying sublexical distributional statistics is within-language modality effects. Indeed, spoken and written language modalities have specific characteristics that have strong impacts on distributional statistics. Certain words are favored in the written language, whereas other words are favored in the oral form. Therefore, the oral lexicon is not identical to the written lexicon. Going back to our previous example, *char* also reveals an important modality difference. Within the same language variety (Quebec French), the statistics of the word *char* are probably vastly different in the written and spoken modalities, because *char* is mainly used in the spoken (and familiar) modality, and only rarely occurs in writing (at least nonmilitary ones). Moreover, whereas written languages often have units that are separated by a blank space, spoken language is a continuous flow of sounds, without silences between words or syllables (Kuhl, 2004). This results in the presence of phenomena such as liaisons between the pronounced words, epenthesis (adding phones), and elision (removing phones) during oral discourse, which accounts for a large amount of the variability between different speakers' productions and results in syllables with different structures. For example, the schwa [ə] in the French word *petit* ("small") is typically removed in familiar contexts, resulting in distinct syllables for informal oral (CCV: [pti] "p'tit"), as compared to formal oral or written French (CVCV: [pə-ti] "petit"). Since sublexical units and their distributional properties are specific—at least in part—to each language and each modality, it is necessary to have access to language-specific and modality-specific corpora in order to be able to fully characterize distributional statistics and to study their impacts on spoken language use.

The objective of this project was twofold: (1) to collect a large corpus of oral language (SyllabO+) from French speakers in Quebec,<sup>3</sup> and (2) from this corpus, to create two sublexical databases documenting the use of syllables and phones as a function of speakers' ages, sexes, and communication contexts (formal vs. informal spoken language).

To the best of our knowledge (see Supplemental Material 1), no database of spoken Quebec French syllables and phones currently exists. A few databases or corpora exist for the French language, notably Lexique 3 (New, Pallier, Brysbaert, & Ferrand, 2004; New, Pallier, Ferrand, & Matos, 2001), Diphones-fr (New & Spinelli, 2013), InfoSyll (Chetail & Mathey, 2010), Texto4Science (Langlais & Drouin, 2012), QUÉBÉTEXT (Trésor de la langue française au Québec, n.d.), and Phonologie du français contemporain (PFC; Durand, Laks, & Lyche, 2001, 2009). The first database, Lexique 3, was created from a corpus of French texts and film subtitles (France)—which represent a hybrid type of corpus that is largely influenced by the written modality. The two

subsequent databases were created from this same corpus: Diphones-fr from Lexique 3, and InfoSyll from Lexique 2 (texts only). These three tools, respectively, offer information on lexical units, pairs of phonemes, and syllables (orthographic and phonological). The fourth database, Texto4Science, is a corpus of text messages (SMS) of French speakers in Quebec, from which all word forms were extracted, as well as their distributional statistics. Although it portrays Quebec French use, Texto4Science is dedicated to the study of written lexical units occurring in the specific context of text messages. QUÉBÉTEXT is composed of four different written Quebec French corpora, providing no information on spoken Quebec French. The last database, PFC, provides recordings and annotated transcriptions of a large number of speakers (>400) from different French-speaking countries. However, PFC does not offer distributional statistics on sublexical units contained in the recordings, and only contains a limited number of speakers from Québec (<20). In sum, all of these databases differ from the present one (SyllabO+) in a number of ways: (1) geographical variety—that is, they do not focus on Quebec French (Lexique3, Diphones-fr, Infosyll, and, for the most part, PFC); (2) modality—that is, they do not focus on spontaneous oral language (Lexique 3, Diphones-fr, InfoSyll, Texto4Science, and QUÉBÉTEXT); and (3) types of data provided—that is, they do not provide sublexical statistics (Lexique 3, Texto4Science, QUÉBÉTEXT, and PFC). None of these databases provides statistics of sublexical use as a function of the characteristics of the speaker and the communication context.

Thus, to facilitate and improve research on spoken language across a variety of disciplines (e.g., psycholinguistics, experimental phonetics, cognitive neuroscience of language, and phonology), we created a unique corpus of contemporary spoken Quebec French, extracted over 300,000 syllables (and over 690,000 phones) from it, and computed a large number of distributional statistics. The resulting tool, composed of a syllable database and a phone database, is called *SyllabO+*. In this article, we present an elaboration of the corpus and databases and describe the Web application that provides access to these databases in open-access format. We also present a brief overview of the corpus data extracted from SyllabO+.

## Method

### Corpus elaboration

Speech samples were collected from 184 different speakers (representing over 300,000 syllables). To ensure that the corpus would accurately represent the use of spoken Quebec French (in both standard and colloquial varieties), samples were collected in both formal and informal communication contexts. The formal contexts consisted mainly of interviews,

<sup>3</sup> In Quebec, a province of Canada, 6,231,600 individuals, representing 80 % of the population, are native speakers of French (Statistique Canada, 2011).

lectures, press conferences, and radio or television programs. In these situations, speakers are more aware of the importance to use “proper” speech than in informal contexts, but we made sure to choose samples that still represented spontaneous speech, and not written discourse that was read aloud. These formal samples represent 63 % of the recordings in the corpus, and 53 % of the total syllables. The samples were obtained mainly through online public resources (92 %)—the recordings were all dated between 2000 and 2014. The online public resources<sup>4</sup> that served to provide recordings for the project consisted mainly of radio and television networks. Though they were not recorded in soundproof rooms, these samples offered a recording quality that ensured high intelligibility (e.g., low background noise), which is necessary for accurate transcription. A small portion of the formal samples were recorded by our team at a lecture or a conference (8 %).

The informal samples represent 37 % of the recordings in the corpus and 47 % of the total syllables. Most of the informal recordings were made in our laboratory or at a participant's home (96 %), between the years 2013 and 2014. A few (4 %) were found through online public resources. The samples collected by our team were recorded in a soundproof room using a small Lavalier microphone clipped to the participant's clothing. Participants were asked to select a few topics that they were comfortable discussing with a team member present in the room, and who used these topics as a conversation opener. The conversation was allowed to evolve freely.

Whether in formal or informal contexts, if only one person's speech was of interest (e.g., if the second person was a team member), only the person of interest's speech was transcribed (including during segments in which the speech overlapped). If more than one person's speech was of interest, everything was transcribed (including the speech segments that overlapped), with a separate transcription file for each participant. In both the formal and informal contexts, we ensured that the topics covered in the speech samples were varied, to ensure representativeness. The topics covered in the corpus include work, education, family, languages, trips, sports, health, cooking, entertainment, environment, society, technology, international, politics, economy, sciences, art, and history.

The participants were all native speakers of Quebec French<sup>5</sup> (mean age  $52 \pm 19.7$  years, range 20–97 years), with a mean of  $16 \pm 3.9$  years of education (range 7–27 years), including 95 male and 89 female speakers. The participants were divided into three age groups: 20–45 years (mean  $32 \pm 6.8$  years), 46–70 years (mean  $55 \pm 7.6$  years), and 71–97 years

(mean  $78 \pm 6.4$  years). The sample is described in Table 1. Participants were recruited through a variety of means: written ads posted in as many strategic locations as possible (e.g., supermarkets, coffee shops, drugstores, local hospitals, and websites), e-mails to large groups (e.g., university students and staff, “golden age” groups), presentations in retirement centers, and contacting individuals in our participant database.

## Transcription

The recordings were first transcribed orthographically. The orthographic transcriptions served as a tool to facilitate the phonetic transcriptions. The phonetic transcriptions were conducted by three different students (P.B., A.-M.A., and J.R.) trained in linguistics and phonetics (see the *Syllabification* section for details about the interjudge agreement). A transcription protocol was elaborated to ensure maximal uniformity and a high level of accuracy. Each sound pronounced was transcribed to the corresponding International Phonetic Alphabet (IPA) symbol. Prosodic characteristics—silences, laughs, onomatopoeia, or other prosodic markers—as well as background noise (nonspeech elements) were not transcribed. It is important to note that the transcription was precise but did not take into account speaker or regional variations in pronunciation, since the goal of the project was not to differentiate between fine regional or interpersonal variations within spoken Quebec French, but to create a tool that would provide information on the use and distribution of syllables and phones in spoken Quebec French in general. Indeed, to obtain syllable frequency scores and other statistics that would be generalizable to Quebec French as a whole, it would not be useful to differentiate syllables on the basis of nondistinctive variations. For example, if the phoneme /e/ was pronounced as a diphthong [e<sup>i</sup>], it was still transcribed [e], since the diphthong is only representative of a few regional accents and is not a distinct phoneme in French. The detailed transcription protocol is provided on the website ([www.speechneurolab.ca/syllabo](http://www.speechneurolab.ca/syllabo)), under “Documentation.”

**Table 1** Numbers of syllables transcribed and numbers of speakers (*n*), as a function of age, sex, and communication context (formal, informal)

Age	20–45 years		46–70 years		71–97 years		Total
	Male	Female	Male	Female	Male	Female	
Formal	27,126 <i>n</i> = 12	28,744 <i>n</i> = 12	27,324 <i>n</i> = 11	27,381 <i>n</i> = 10	25,043 <i>n</i> = 11	25,978 <i>n</i> = 12	161,596 <i>n</i> = 68
Informal	25,527 <i>n</i> = 25	25,682 <i>n</i> = 22	25,100 <i>n</i> = 22	25,941 <i>n</i> = 21	25,659 <i>n</i> = 14	14,131 <i>n</i> = 12	142,040 <i>n</i> = 116
Total	52,653 <i>n</i> = 37	54,426 <i>n</i> = 34	52,424 <i>n</i> = 33	53,322 <i>n</i> = 31	50,702 <i>n</i> = 25	40,109 <i>n</i> = 24	303,636 <i>n</i> = 184

<sup>4</sup> The main resources were the following: Radio-Canada, TVA, RDS, TOU.TV, Canal Vie, Assemblée nationale du Québec, and the Montreal and Quebec city official video channels.

<sup>5</sup> They were born in Quebec and reported Quebec French as their native language (i.e., the language was learned at home via parents speaking Quebec French).

## Syllabification

Each phonetic transcription was syllabified; that is, the continuous speech transcriptions were divided into individual syllables. This step was crucial, as it is at this point that the units of interest (the syllables) were obtained. The same three transcribers executed this step. First, a detailed syllabification protocol was established (to consult the detailed protocol, visit [www.speechneurolab.ca/syllabo](http://www.speechneurolab.ca/syllabo), under “Documentation”). The inherent structure of French syllables was used as a guideline for syllabification, whereby consonants at onset were favored rather than consonants at coda (e.g., Brousseau & Nikiema, 2001; Noske, 1982; Paradis, 1993). Within words, the least marked syllables (CV, CVC, and CCV) were preferred to other syllables. For example, the French word *professeur* (“teacher”) would be transcribed [pr fɛsœr] and not \*[pr fɛsœr].<sup>6</sup> Elided segments as well as liaisons were taken into account in the syllabification process. For example, *sur le pouce* (“on the go”) would be transcribed [syl pus] if such was the pronunciation, and elided segments would not be reconstructed into “expected” pronunciation \*[syr læ pus]. Likewise, if sounds were added for liaisons between words, they would be transcribed as pronounced; *les amis* (“the friends”) would be transcribed [lɛ zami] according to the speaker’s pronunciation.

Consistency between transcribers was assessed by calculating an interjudge agreement between all transcribers on a subset of the recordings at the beginning of the project, during the ongoing project, and at the end. To establish the agreement, the three transcribers produced a transcription of 20,775 syllables from the same 120 min of speech recordings (representing 18 different speakers: ten in formal context, eight in informal context). These 20,775 syllables represent 7 % of the entire corpus (303,636 syllables), and about 10 % of all speakers. The interjudge agreement was established between the transcriptions of the three transcribers at once (one recording), or between two of the three transcribers at once for the same recording (17 recordings). Specifically, the transcriptions of Transcriber 1 were compared to those of Transcriber 2, and so on for Transcribers 2 and 3 and Transcribers 3 and 1, so that the transcriptions from all three transcribers were ultimately compared to each other. The interjudge agreement was calculated from these common transcriptions, in which each dissimilar IPA symbol or dissimilar syllable counted as an error. The percentage of errors was calculated on the total number of symbols contained in the transcription. Interjudge agreements of 90 % or more were obtained for the informal speech contexts, and 94 % or more were obtained for the formal speech contexts. The average interjudge agreement was 94.90 ± 2.17 %. Moreover, the interjudge agreement remained constant

<sup>6</sup> Note that in this paragraph, the asterisks represent inaccurate forms, thus differentiating between the preferred and the erroneous form.

throughout the project. At the beginning of the project, 14 recordings (13,131 syllables) were evaluated, with an average agreement of 95.68 %. One recording was compared during the ongoing project (1,382 syllables), with an average agreement between all three transcribers of 92.53 %. Finally, three recordings (6,262 syllables) were compared at the end, with an average agreement of 94.15 %. The dissimilarities between the transcribers were resolved by agreeing on a specific phone or syllable and by clarifying the transcription protocol.

## Creation of syllable and phone databases (SyllabO+)

The syllabified transcriptions were saved as annotated and marked-up XML files. All metadata (anonymized information about the speaker or the recording) were saved in another XML file and were linked to each individual transcription by a reference number. Extracting statistical information from these XML files was done by means of a Python script, which enabled automatic calculations of a number of distributional statistics. The extracted statistical information was then organized in tables, which constitute the databases.

The syllable and phone databases were integrated into an open-access Web application, developed by a team of expert programmers. The purpose of the Web application was to provide access to both the databases and the corpus to researchers and students from a variety of disciplines, as well as to knowledge users, such as speech–language pathologists and language teachers, thereby maximizing the use and impact of SyllabO+.

## Description of the syllable database

The syllable database consists of four different data tables: unique syllables with related data and statistics, and syllable collocations, which include pairs of syllables with related data and statistics, groups of three syllables with related data and statistics, and groups of four syllables with related data and statistics. A description of the database tables, with definitions and detailed description of calculations is available on our [www.speechneurolab.ca/syllabo](http://www.speechneurolab.ca/syllabo), under “Documentation.”

The complete database, or a specific subset of the database resulting from specific query options, can be downloaded from the Web. The entire corpus is available upon e-mail requests to the corresponding author. The following parameters can be used individually or in combination: context of communication (formal, informal), age (range), and sex of the speakers. The files can be downloaded in CSV (comma-separated value) format, which is a way of storing tabular data in plain text—in this case, UTF-8 text. These files can be opened and used with spreadsheet software or database software (e.g., LibreOffice, MySQL). The recommended method to open the CSV files is with LibreOffice, a free and open-source software application ([www.libreoffice.org](http://www.libreoffice.org)). The main

reason for this recommendation is that Microsoft Excel does not handle the IPA characters correctly, whereas LibreOffice does (OpenOffice and Google Sheets online do, as well). In LibreOffice, select “Unicode (UTF-8)” as the character set when opening the CSV file. For users who prefer to use other software or who have issues with special-character display, we provide two options: Excel files in which the IPA characters have already been encoded, and are therefore readable, or CSV files in which all IPA characters have been converted to the Speech Assessment Methods Phonetic Alphabet (SAMPA; consisting only of ASCII characters), which will avoid any issue with character display in Excel or any other software. For an extensive description of how to use SyllabO+, refer to the user manual, available on our website at [www.speechneurolab.ca/syllabo](http://www.speechneurolab.ca/syllabo), under “Documentation.”

### Description of the phone database

The phone database was elaborated from the same corpus from which the individual sounds were extracted. The database consists of four different data tables: unique phones with related data and statistics, and phone collocations, which included diphones with related data and statistics, triphones with related data and statistics, and tetraphones with related data and statistics. The data provided in the tables include phone structure, frequency (raw and percentage), percentile rank, transition probabilities, and mutual information. The information provided for phone collocations differs from the syllabic information, since the phone collocations were extracted irrespective of word or syllable boundaries (i.e., through an entire speech recording). Syllables, in contrast, are either words or part words. The phone database is available in the same Web application as the syllable database.

### General description: syllable database

The corpus contains a total of 303,636 syllables, including 48 different phones, composed of 22 consonants ([p] [t] [k] [b] [d] [g] [f] [s] [ʃ] [v] [z] [ʒ] [m] [n] [ŋ] [l] [r] [ɹ] [ɹ̥] [ð] [h]<sup>7</sup> [x]<sup>8</sup>), 23 vowels ([i] [y] [u] [e] [ø] [o] [ə] [ɛ] [œ] [ɔ] [a] [ɑ] [ɛ̃] [ã] [ɔ̃] [œ̃] [ʌ] [ɔ̃] [ɜ̃] [æ̃] [ɪ̃] [ỹ] [ʊ̃]), and three semi-vowels ([w] [j] [ɥ]), which combine to form 5,213 different

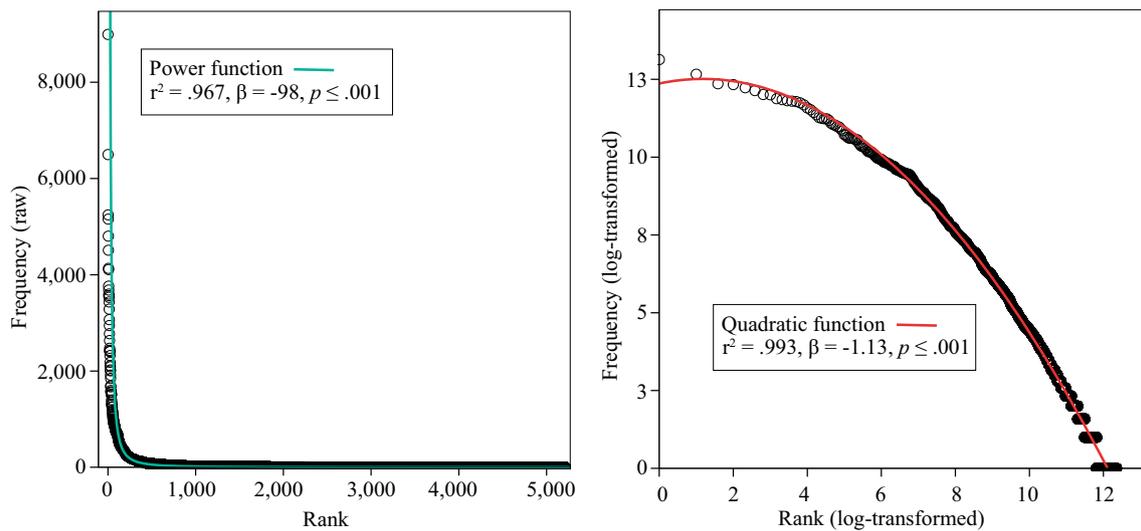
syllables. To document the distributions of syllables and phones, here we present two different kinds of frequencies: (1) the token frequency (i.e., the total numbers of occurrences of all units [syllables or phones]), and (2) the type frequency (i.e., the numbers of different units, not taking into account how many times each unit occurs).

Of all the 5,213 different syllables, a small subset are used extremely frequently. Indeed, the 5 % most frequently used syllables represent 78 % of the corpus (total syllables pronounced). The same is true for pairs of syllables, for which the 5 % most frequently used pairs represent 48 % of the corpus (total pairs pronounced). Tables of the most frequent syllables and the most frequent pairs can be found directly on SyllabO+. Figure 1 illustrates the distribution of the syllables in absolute (rank) frequencies and on a logarithmic scale (with base 2). The raw frequencies range from 1 to 8,994. A large number of syllables (1,024 out of a total 5,213 different syllables, representing 19.64 %) have a frequency of 1. The mean of the distribution is 58.246, and the median is 3. As can be seen from the figure, the distribution is asymmetrical, with a pronounced right skew. To characterize these distributions, seven different mathematical models were fitted to both the raw and the log-transformed data (linear, inverse, quadratic, cubic, power, exponential, and growth) using SPSS Statistics version 23 (IBM). For the raw frequency, the best fit was found using a power model ( $r^2 = .967$ ,  $\beta = -98$ ,  $p \leq .001$ , where  $y = 2,769,892.4 \times x^{-1.74}$ ), whereas for the log-transformed data, the best fit was found using a quadratic model ( $r^2 = .993$ ,  $\beta = -1.13$ ,  $p \leq .001$ , where  $y = 12.37 + 0.25x + -0.1x^2$ ). These results are provided in Fig. 1.

The number of different syllable structures and each structure's prevalence were also investigated. Within the 5 % most frequent syllables (representing 260 different syllables), only 12 syllable structures were found. Figure 2 describes the 5 % most frequent syllables in terms of their structures, taking into account the number of occurrences of every syllable—that is, how many times each syllable is repeated (token frequency) and the number of different syllables within each syllabic structure (type frequency) (e.g., the corpus contains 138 different CV syllables, which account for 53 % of all utterances). Supplemental Material 2 presents all of the syllabic structures across the entire corpus in the same way. Interestingly, for the 5 % most frequent syllables (Fig. 2), the most frequent syllabic structures (CV) also accounts for the largest proportion of the different syllabic structures within this subset (out of a total 260 different structures). In contrast, in the entire corpus (Supplemental Material 2), only 5 % of all different syllables (out of a total 5,213 different structures) have a CV structure, but 52 % of all pronounced syllables have a CV structure. This means that there are a limited number of different CV syllables, but these CV syllables are used extremely frequently and account for more than half of the corpus (i.e., the total number of syllables pronounced).

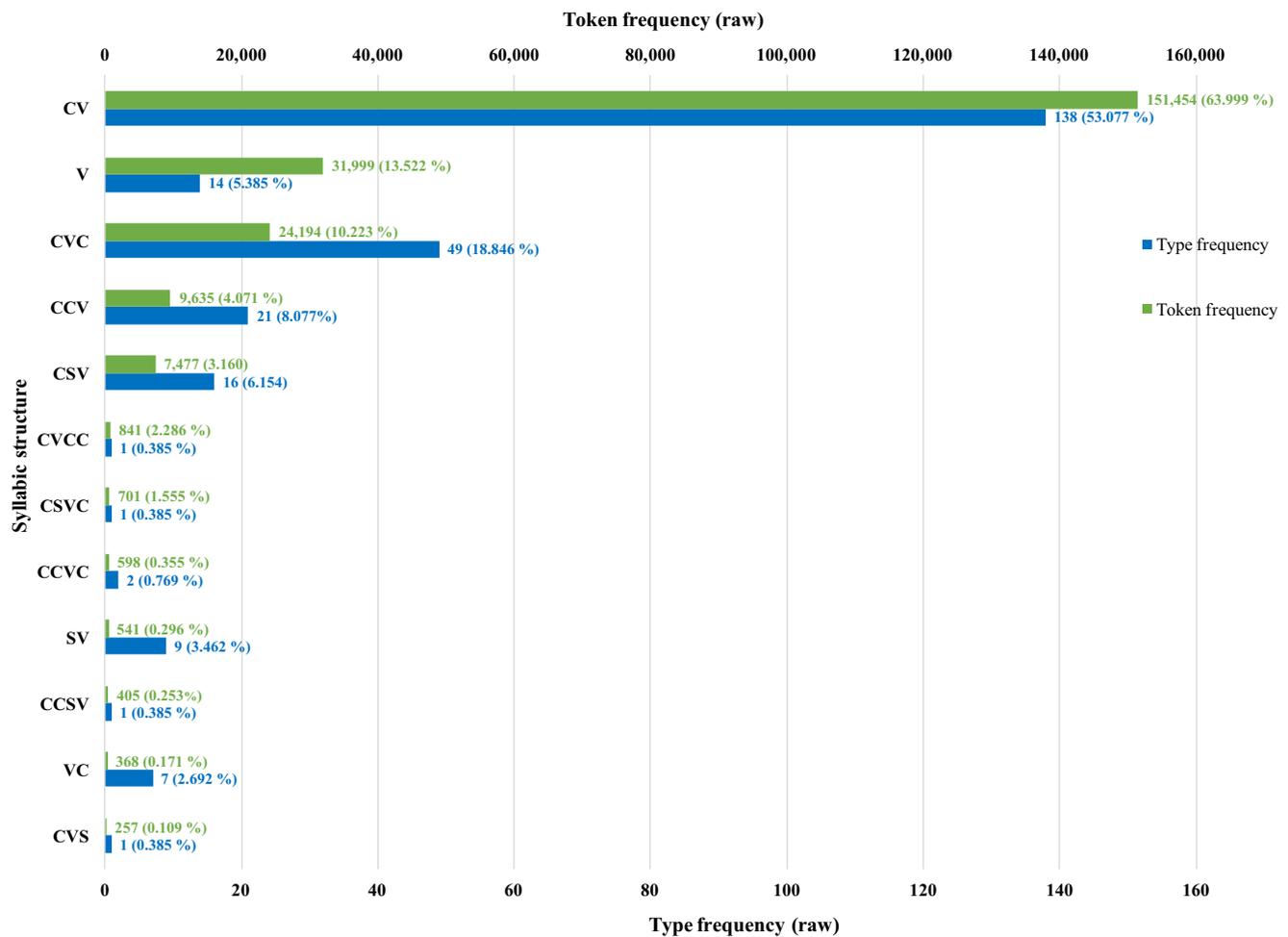
<sup>7</sup> These phones are used only for English pronunciations. Note that the proportion of English-pronounced syllables (containing one or more typically English sounds) in the corpus is 458 out of 303,636 syllables (0.1508 %). It is possible to filter out these syllables (or phones) in the database tables by excluding any that contain the English-specific consonants and vowels used in the transcriptions, which are [ɹ̥ ð h] and [ʌ ɔ̃ ɜ̃ æ̃ ɪ̃ ỹ ʊ̃].

<sup>8</sup> [x] is used only for Spanish pronunciations. Note that the proportion of Spanish-pronounced syllables (containing the typically Spanish sound [x]) in the corpus is 5 out of 303,636 syllables (0.0016 %). It is possible to filter out these syllables (or phones) in the database tables by excluding any that contain the Spanish-specific consonant [x].



**Fig. 1** Distribution of syllable frequencies. (a) Rank frequencies, which associate any given frequency to the syllable in the corresponding rank position and enable the representation of the frequency distribution (i.e., the syllable frequencies in descending order), and (b) log–log plot of

syllable frequencies (log base 2). For each graph, the mathematical model that best fitted the distribution is illustrated, and the fit is characterized in terms of  $r^2$ , slope (standardized beta coefficient), and statistical significance ( $p$ ). The specific equations can be found in the text



**Fig. 2** Token frequency and type frequency for each syllabic structure (extracted from the 5 % most frequent syllables), in raw frequency and percentage frequency (%)

In the corpus, the syllabic structure that accounts for the largest number of different syllables is CVC (27 %; Supplemental Material 2). The CVC structure is thus very “prolific” for generating different syllables, in the sense that it allows for many different combinations (hypothetically, 22 consonants × 23 vowels × 22 consonants), while remaining a relatively simple structure (it does not contain any consonant clusters and contains only three phones). The CV structure is less prolific than the CVC structure in generating different syllables (its combinations are limited to, hypothetically, 22 consonant × 23 vowels, which may account for this difference). However, CV syllables have, on average, higher frequencies than the CVC syllables.

Importantly, the entire corpus contains a much larger number of structures than the ones present within the 5 % most frequent syllables: 52 different structures (Supplemental Material 2), as compared to only 12 (Fig. 2). This shows that there is a large number of possibilities when it comes to syllabic structures in French, but that many of these structures are seldom used, probably because of their complexity (e.g., the presence of one or more consonant clusters, such as in CCVCC<sup>9</sup>; Supplemental Material 2).

Finally, we also explored the distribution of syllables in the corpus as a function of their syllabic structures, in terms of the presence of at least one consonant cluster. Figure 3 presents a summary of the distribution of syllables containing a consonant cluster, as a function of the position of the cluster within the syllable (onset, coda, or both). The detailed information about the distribution of complex syllables can be found in Supplemental Material 3. This analysis shows that complex syllable structures represent ~50 % of the structures present in the corpus, but account for only ~11 % of all syllables (token frequency). The syllables with a consonant cluster at onset are far more frequent in the corpus than the two other types, in terms of both total frequency (respectively, 9.7 % of all syllables, vs. 1.2 % and 0.1 %) and the number of different syllables (respectively, 41.4 % of all different syllables, vs. 8.2 % and 2.0 %). The most frequent structure with a consonant cluster at the onset is CCV (70.4 %); the most frequent structure with a consonant cluster at the coda is CVCC (89.5 %); and the most frequent structure with a consonant cluster at both onset and coda is CCVCC (94.8 %).

<sup>9</sup> Syllable structures in French can be described as increasing in complexity when increasing in (1) the number of consonant clusters and (2) the number of consonants per consonant cluster. This is consistent with Noske (1982), where the increase in markedness of a syllable structure is linked to consonant clusters (p. 270):

Onset	Rime	Markedness
C	V	0
∅	∅	1
CC	VC	2
CCC	VCC	3
C <sub>1</sub> ... C <sub>n</sub>	VC <sub>1</sub> ... VC <sub>n-1</sub>	"

### General description: phone database

The corpus contains a total of 692,707 phones, including 48 different phones (see the general description of the corpus). The number of different consonants is nearly identical to the number of different vowels: respectively, 22 and 23. Moreover, the total frequencies are also similar, with 363,062 consonants (52.41 % of the corpus), 303,635 vowels (43.83 %), and 25,934 semivowels (3.74 %).

Figure 4 illustrates the distribution of phone frequencies and presents each phone with its structure (C, V, or S).

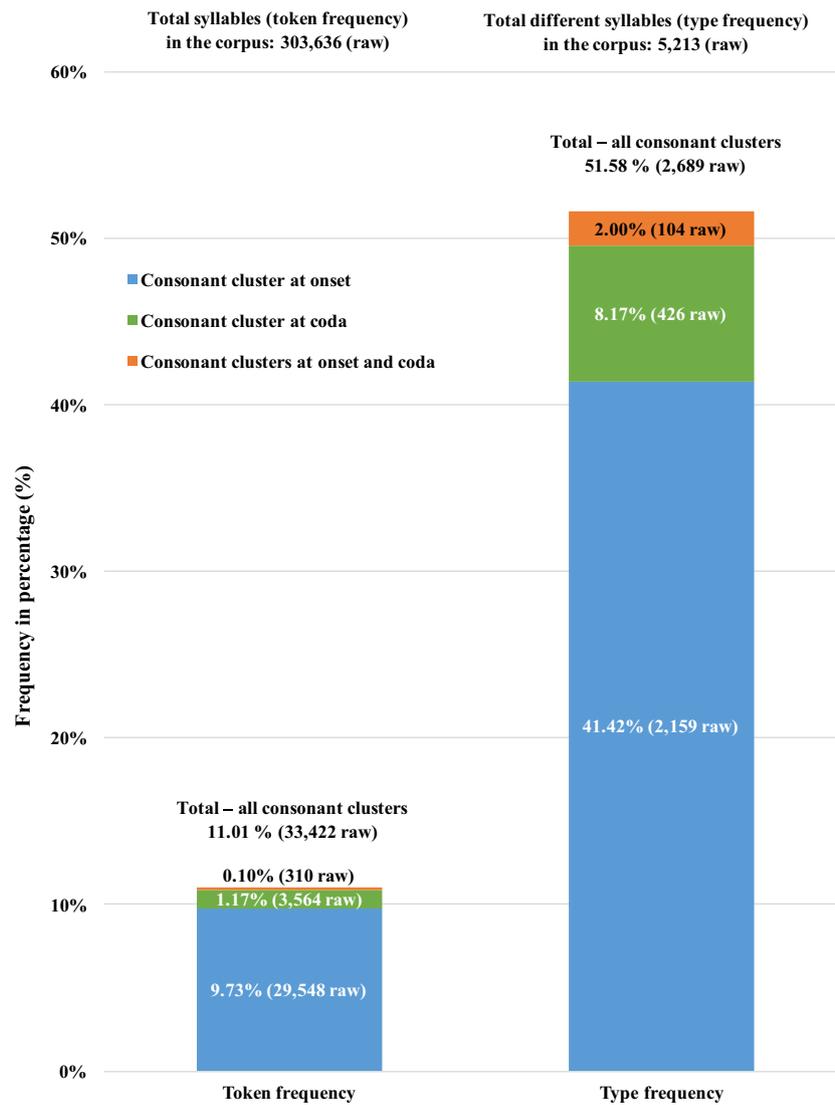
### Limitations

The main limitation of SyllabO+ is its size (303,636 syllables/692,707 phones, but 184 different speakers), which may appear small, especially when compared to certain lexical databases (extracted from written corpora) that comprise millions of occurrences. However, given the substantial amount of time and work required for the conversion of speech samples to transcribed syllables in phonetic alphabet and the calculation of their distributional statistics, and given the fact that resources are not unlimited, SyllabO+'s size is not only impressive, but also a fair representation of the spoken language in native adult speakers in Quebec (all age groups). SyllabO+ is, by nature, a specialized resource, representing Quebec French in a specific modality (spoken) while focusing on sublexical units. Given the large amount of information provided in SyllabO+ (syllable and phone structures, frequencies, transition probabilities, and mutual information) and because of the uniqueness and versatility of this tool, we believe it will serve a number of research and clinical purposes (see the Conclusion for a discussion of some potential uses of SyllabO+).

### Conclusion

SyllabO+, a multispeaker corpus of spoken Quebec French, addresses the need for a tool focusing on spoken French in Quebec. It will allow researchers to study the spontaneous use of French in Quebec in younger and older male and female adult speakers in both formal and informal communication contexts. The next phase of the project will focus on developing an additional word corpus. This will be done by creating a lexical table containing all word occurrences from the corpus and their associated distributional values. This enrichment will turn SyllabO+ into one of the most comprehensive linguistic resources on oral language.

SyllabO+ is an unparalleled resource: not only is it the first sublexical database on spoken Quebec French, but is also one of the richest linguistic databases in terms of the



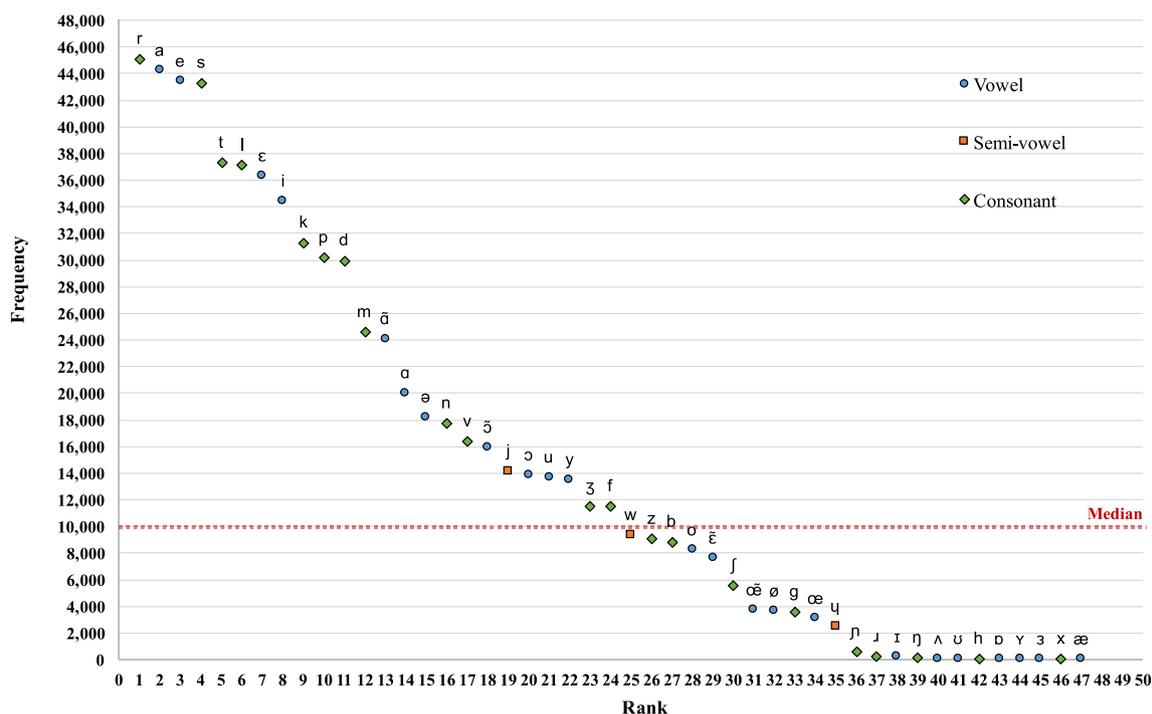
**Fig. 3** Token frequencies and type frequencies for complex syllabic structures—that is, those including at least one consonant cluster—in percentage frequency (%) and raw frequency

amount of statistical information it provides (e.g., transition probabilities and mutual information), all of which are available as a function of speakers' age, sex, and communication context. Indeed, unlike most available linguistic databases, SyllabO+ allows the extraction of multiple specific subdatabases by using specific search parameters (sex, age, and context). Users can thus conduct a large number of analyses according to their interests, whether these be context-related comparisons (e.g., comparing the use of *char* in formal and informal spoken French) or comparative studies by age group.

SyllabO+'s applications are manifold. Researchers in the fields of cognitive neuroscience of language, psycholinguistics, neurolinguistics, and experimental psychology can use it to create stimuli representative of the spoken language—both sublexical (e.g., syllable) and lexical (word) stimuli—controlled for normalized frequency, transition probabilities, or

mutual information. Given the known effects of distributional statistics on speech perception and production, behaviorally and at the neural level (Carreiras, Mechelli, & Price, 2006; Carreiras & Perea, 2004; Cibelli, Leonard, Johnson, & Chang, 2015; Deschamps et al., 2016; Karuza et al., 2013; Leonard et al., 2015; Newport & Aslin, 2004; Pelucchi et al., 2009a, 2009b; Peña et al., 2002; Saffran et al., 1996; Saffran, Johnson, Aslin, & Newport, 1999; Tremblay et al., 2012; Tremblay et al., 2016; Vitevitch, 2003; Vitevitch & Luce, 1998; Vitevitch et al., 1997, 1999) it is important to control for these effects in order to avoid confounds that can mask other effects of interest, such as articulatory or phonological complexity. However, such important experimental control is only possible when databases providing this information exist, which was not the case for Quebec spoken French.

In addition to its use as a key experimental control, the distributional statistics provided in SyllabO+ can be used



**Fig. 4** Distribution of phone frequencies (raw)

to study, using behavioral and brain-imaging approaches, the impact of distributional frequencies on a variety of language production and comprehension tasks, by manipulating, rather than controlling, sublexical properties such as syllable complexity and frequency. This opens a wealth of research avenues for researchers interested in understanding how the human brain processes distributional information and how this information contributes to language comprehension.

Furthermore, SyllabO+ will also have multiple uses in linguistics. Indeed, linguists will be able to use it to explore sublexical phenomena, notably in terms of differences and similarities with documented lexical phenomena (both spoken and written). For example, it is interesting to note that the distribution pattern of the syllables found in our corpus, whereby a small subset of units are produced extremely frequently, is analogous to lexical distributions (e.g., the “kernel lexicon”) that have been reported (Cancho & Solé, 2001). This suggests that a small number of very frequent syllables are used to create a small number of extremely common words. Additional analyses will be needed to examine in more detail the relationship between syllable and word usage in spoken language. At present, direct comparisons, and thus strong conclusions, are not warranted, given that we have not yet extracted a list of all words from our corpus.

Moreover, experts in phonetics and phonology will be able to use SyllabO+ to study spontaneous Quebec oral French and its sublexical phenomena (e.g., liaisons,

or prolificacy of the different syllabic structures), including age-related, gender-related, and context-related phenomena, as well as to observe phonotactic regularities appearing in the corpus. SyllabO+ also has applications in the field of comparative linguistics, since it will allow researchers to compare spoken language use in Quebec to language use in other French-speaking countries for which distributional statistics are available, or even across different languages—for instance, to compare the use of different syllabic structures across languages. We hope that SyllabO+ will be the basis for a number of new studies, whether descriptive (detailing the particularities of Quebec French) or comparative (exploring linguistic similarities and differences across languages, modalities, etc.), originating from as many groups of researchers as possible.

We also expect SyllabO+ to be useful to knowledge users beyond academia, such as speech–language pathologists and language teachers, who can use SyllabO+ to elaborate targeted and ecological intervention plans based on the actual use of syllables of different phonological complexities at different ages. In particular, the frequency of use of the different syllabic structures (e.g., CV, CCV, CCVC) and their prolificacy will be useful to clinicians. For instance, a clinician may want to know the syllable structures that are most common in spoken language, since these are the most important to recover as part of an intervention. Moreover, understanding whether spoken language uses

change over time, and whether there are gender-related differences in spoken language use, will provide a framework against which one can compare the oral productions of individuals with speech perception and production deficits.

Finally, SyllabO+ may serve computer science purposes by enabling the creation of new tools for natural language processing. Orthographic data could eventually be aligned to the phonetic data to develop algorithms that automatically translate written language to an oral representation. Numerous technologies use algorithms that create automata<sup>10</sup> (e.g., finite-state automata) from phonological rules (Jurafsky & Martin, 2000). One of the main uses is vocal synthesis (text to speech), which converts a text into a phonetic representation that is then “pronounced” (actualized in acoustical waves) by a speech synthesizer. This type of processing can be achieved through a machine-learning system, in which a model is generated automatically from a set of data (e.g., by “learning” phonological rules). SyllabO+’s data could therefore serve as a starting point for such computational processing, focusing on representing the Quebec French language. Moreover, the data from SyllabO+ can be useful to computational linguists and data scientists interested in the global structure of language, since they represent a unique and novel source of spoken data (especially in a field in which the majority of the data analyzed is written).

In sum, we hope that SyllabO+ will be widely used to study sublexical phenomena in different fields of research, ranging from cognitive neuroscience to computational linguistics and speech–language pathology.

**Acknowledgments** This project was supported by an Insight Development Grant from the Social Sciences and Humanities Research Council of Canada to P.T. (Grant No. 430-2013-1084), and by an infrastructure grant from the Canada Foundation for Innovation (CFI), also to P.T. (Leaders Opportunity Fund 31408). P.T. holds a Career Award from the “Fonds de Recherche du Québec–Santé” (Grant No. 27170). P.B. was supported by a scholarship from the Natural Sciences and Engineering

Research Council of Canada and by a Wilbrod Bhéner scholarship from the Faculté de Médecine de l’Université Laval. We thank Claudie Ouellet for her help with participant recruitment and testing, Alexis Piéplu for his help with the development of our website, the company Savoir-Faire Linux for the development of the Web application, and all the speakers.

## References

Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, *88*, 148–159. doi:10.1016/S0093-934X(03)00296-7

<sup>10</sup> Finite-state automata consist in mathematical models of computation. They are used to develop computer programs and sequential logic circuits.

- Akamatsu, T. (1967). Quelques statistiques sur la fréquence d’utilisation des voyelles nasales françaises. *La Linguistique*, *3*, 75–80.
- Brousseau, A.-M., & Nikiema, E. (2001). *Phonologie et morphologie du français*. Saint-Laurent, Québec: Fides.
- Cancho, R. F. I., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B*, *268*, 2261–2265. doi:10.1098/rspb.2001.1800
- Canepari, L. (2005). *A handbook of pronunciation: English, Italian, French, German, Spanish, Portuguese, Russian, Arabic, Hindi, Chinese, Japanese, Esperanto*. München: Lincom Europa.
- Carreiras, M., Mechelli, A., & Price, C. J. (2006). Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human Brain Mapping*, *27*, 963–972. doi:10.1002/hbm.20236
- Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. *Brain and Language*, *90*, 393–400. doi:10.1016/j.bandl.2003.12.003
- Chetail, F., & Mathey, S. (2010). InfoSyll: A syllabary providing statistical information on phonological and orthographic syllables. *Journal of Psycholinguistic Research*, *39*, 485–504. doi:10.1007/s10936-009-9146-y
- Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, *99*, 205–235. doi:10.1016/j.cognition.2005.01.009
- Cibelli, E. S., Leonard, M. K., Johnson, K., & Chang, E. F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and Language*, *147*, 66–75. doi:10.1016/j.bandl.2015.05.005
- Ciolac, A. (2010). Aperçu des conceptions portant sur le français québécois. *Revue Roumaine de Linguistique*, *55*, 271–291.
- Deschamps, I., Hasson, U., & Tremblay, P. (2016). The structural correlates of statistical information processing during speech perception. *PLoS ONE*, *11*, e0149375. doi:10.1371/journal.pone.0149375
- Durand, J., Laks, B., & Lyche, C. (2001). La phonologie du français contemporain: Usages, variétés et structure. In C. Pusch & W. Raible (Eds.), *Romanische Korpuslinguistik: Korpora und gesprochene Sprache/Romance corpus linguistics—Corpora and spoken language* (pp. 93–106). Tübingen, Germany: Gunter Narr Verlag.
- Durand, J., Laks, B., & Lyche, C. (2009). Le projet PFC: Une source de données primaires structurées. In J. Durand, B. Laks, & C. Lyche (Eds.), *Phonologie, variation et accents du français* (pp. 19–61). Paris, France: Hermès.
- Gess, R., Lyche, C., & Meisenburg, T. (2012). *Phonological variation in French: Illustrations from three continents*. Amsterdam, The Netherlands: Benjamins.
- Goslin, J., Galluzzi, C., & Romani, C. (2013). PhonItalia: A phonological lexicon for Italian. *Behavior Research Methods*, *46*, 872–886. doi:10.3758/s13428-013-0400-8
- Goyet, L., Nishibayashi, L.-L., & Nazzi, T. (2013). Early syllabic segmentation of fluent speech by infants acquiring French. *PLoS ONE*, *8*, e79646. doi:10.1371/journal.pone.0079646
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.
- Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*, 46–54. doi:10.1016/j.bandl.2012.11.007
- Klein, J. R., & Rossari, C. (2003). Set phrases and variation in the French of Belgium, France, Quebec, and Switzerland. *Linguisticae Investigationes*, *26*, 203–214.
- Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843. doi:10.1038/nrn1533

- Langlais, P., & Drouin, P. (2012). Texto4Science: A Quebec French database of annotated text messages. *Linguisticae Investigationes*, 35, 237–259.
- Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience*, 35, 7203–7214. doi:10.1523/JNEUROSCI.4100-14.2015
- Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, 3, 223–232. doi:10.1016/S1364-6613(99)01319-4
- Levelt, C. C., Schiller, N. O., & Levelt, W. J. (2000). The acquisition of syllable types. *Language Acquisition*, 8, 237–264. doi:10.1207/S15327817LA0803\_2
- Lleó, C., & Prinz, M. (1996). Consonant clusters in child phonology and the directionality of syllable structure assignment. *Journal of Child Language*, 23, 31–56. doi:10.1017/S0305000900010084
- Martin, P., Beaudoin-Begin, A.-M., Goulet, M.-J., & Roy, J.-P. (2001). Les voyelles nasales en français du Québec. *La Linguistique*, 37, 49–70.
- McLeod, S., van Doorn, J., & Reed, V. A. (2001). Normal acquisition of consonant clusters. *American Journal of Speech-Language Pathology*, 10, 99–110.
- Mercier, L. (2002). Le français, une langue qui varie selon les contextes. In *Français, une langue à apprivoiser* (pp. 41–60). Sainte-Foy, Québec: Presses de l'Université Laval.
- Moreau, M. L. (1997). *Sociolinguistique: les concepts de base*. Mardaga. Retrieved from <https://books.google.fr/books?id=rLG73PRRKd4C>
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36, 516–524. doi:10.3758/BF03195598
- New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE<sup>1</sup>. *L'Année Psychologique*, 101, 447–462.
- New, B., & Spinelli, E. (2013). Diphones-fr: A French database of diphone positional frequency. *Behavior Research Methods*, 45, 758–764. doi:10.3758/s13428-012-0285-y
- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127–162. doi:10.1016/S0010-0285(03)00128-2
- Noske, R. (1982). Syllabification and syllable changing rules in French. In H. V. D. Hulst & N. Smith (Eds.), *The structure of phonological representations* (Vol. 2, pp. 257–310). Dordrecht, The Netherlands: Foris.
- Paradis, C. (1993). Phonologie générative multilinéaire. In *Tendances actuelles en linguistique générale* (pp. 11–47). Neuchâtel, France: Delachaux et Niestlé.
- Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113, 244–247. doi:10.1016/j.cognition.2009.07.011
- Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80, 674–685. doi:10.1111/j.1467-8624.2009.01290.x
- Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298, 604–607. doi:10.1126/science.1072901
- Podesva, R. J., & Sharma, D. (2014). *Research methods in linguistics*. Cambridge, UK: Cambridge University Press.
- Romani, C., Galluzzi, C., Bureca, I., & Olson, A. (2011). Effects of syllable structure in aphasic errors: Implications for a new model of speech production. *Cognitive Psychology*, 62, 151–192. doi:10.1016/j.cogpsych.2010.08.001
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928. doi:10.1126/science.274.5294.1926
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52. doi:10.1016/S0010-0277(98)00075-4
- Schiel, F. (2010). BAStat: New statistical resources at the Bavarian Archive for Speech Signals. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, ... D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1069–1076). Luxembourg City: European Language Resources Association.
- Staiger, A., & Ziegler, W. (2008). Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech. *Aphasiology*, 22, 1201–1215. doi:10.1080/02687030701820584
- Statistique Canada. (2011). *Tableau 2 Effectif et proportion de la population ayant déclaré le français selon la caractéristique linguistique, Québec, 2006 et 2011*. Retrieved March 10, 2016, from [https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/2011003/tbl/tbl3\\_1-2-fra.cfm](https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/2011003/tbl/tbl3_1-2-fra.cfm)
- Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, 10, 21.
- Tremblay, P., Baroni, M., & Hasson, U. (2012). Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure. *NeuroImage*, 66, 318–332. doi:10.1016/j.neuroimage.2012.10.055
- Tremblay, P., Deschamps, I., Baroni, M., & Hasson, U. (2016). Neural sensitivity to syllable frequency and mutual information in speech perception and production. *NeuroImage*, 136, 106–121. doi:10.1016/j.neuroimage.2016.05.018
- Trésor de la langue française au Québec. (n.d.). *QUÉBÉTEXT*. Retrieved March 10, 2016, from [www.tlfq.ulaval.ca/quebetext/default.asp](http://www.tlfq.ulaval.ca/quebetext/default.asp)
- Vaden, K. I., Kuchinsky, S. E., Keren, N. I., Harris, K. C., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2011). Inferior frontal sensitivity to common speech sounds is amplified by increasing word intelligibility. *Neuropsychologia*, 49, 3563–3572. doi:10.1016/j.neuropsychologia.2011.09.008
- Vaden, K. I., Piquado, T., & Hickok, G. (2011). Sublexical properties of spoken words modulate activity in Broca's area but not superior temporal cortex: Implications for models of speech recognition. *Journal of Cognitive Neuroscience*, 23, 2665–2674. doi:10.1162/jocn.2011.21620
- Vajta, K. (2012). Autant en emporte le vin, ou: De l'importance des voyelles nasales. *Moderna Språk*, 106, 145–156.
- Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, 17, 487–499. doi:10.1080/0269920031000107541
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, 9, 325–329. doi:10.1111/1467-9280.00064
- Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40, 47–62. doi:10.1177/002383099704000103
- Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68, 306–311. doi:10.1006/brln.1999.2116