Behavior Research Methods (2025) 57:47
https://doi.org/10.3758/s13428-024-02582-2

ORIGINAL MANUSCRIPT

Expansion of the SyllabO+ corpus and database: Words, lemmas, and morphology

Noémie Auclair-Ouellet[1] · Alexandra Lavoie[2,3] · Pascale Bédard[2] · Alexandra Barbeau-Morrison[1] · Patrick Drouin[4] · Pascale Tremblay[2,3]

Accepted: 1 December 2024

Abstract

Having a detailed description of the psycholinguistic properties of a language is essential for conducting well-controlled language experiments. However, there is a paucity of databases for some languages and regional varieties, including Québec French. The SyllabO+ corpus was created to provide a complete phonological and syllabic analysis of a corpus of spoken Québec French. In the present study, the corpus was expanded with 41 additional speakers, bringing the total to 225. The analysis was also expanded to include three new databases: unique words, lemmas, and morphemes (inflectional, derivational, and compounds). Next, the internal structure of unique words was analyzed to identify roots, inflectional markers, and affixes, as well as the components of compounds. Additionally, a group of 441 speakers of Québec French provided semantic transparency ratings for 3764 derived words. Results from the semantic transparency judgment study show broad inter-individual variability for words of medium transparency. No influence of sociodemographic variables was found. Transparency ratings are coherent with studies showing the greater transparency of suffixed words compared to prefixed words. Results for participants who speak French as a second language support the association between second-language proficiency and morphological processing.

Keywords Corpus · Oral language · Distributional statistics · Words · Lemmas · Morphology · Inflectional morphology · Derivational morphology · Composition

Introduction

Words are meaning units encapsulated in a phonological and, in many languages, an orthographic form. The semantic, phonological, and, where applicable, orthographic properties of words collectively shape a language's psycholinguistic characteristics, influencing how speakers process, learn, and use language to communicate. Access to detailed descriptions of these properties is invaluable for investigating typical language processing as well as developmental and acquired language disorders.

SyllabO+ was initially created to provide a complete phonological and syllabic analysis of a corpus of spoken Québec French (Bédard et al., 2017), with a focus on statistics about sublexical units (e.g., syllables and syllable strings, phonemes). Indeed, SyllabO includes the frequency of use of each unit, as well as its transition probabilities—the probability of transitioning from one syllable to another in a single step—and other relational statistics. Access to statistical information, including transition probabilities, is crucial for learning how to divide continuous speech into syllables and words, particularly in early childhood where word boundaries may not be apparent. Research has shown that both infants and adults are responsive to transition probabilities in speech. This sensitivity allows for the prediction of forthcoming syllables and words (Newport & Aslin, 2004; Pelucchi et al., 2009a, b; Saffran et al., 1996, 1999). For example, if a certain syllable, say "/muv/", is consistently followed by syllable "/mã/" (in the syllabO+ database, "/

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

[1] School of Communication Sciences and Disorders, McGill University, Montréal, QC, Canada

[2] Centre de recherche CERVO, Québec City, QC, Canada

[3] École des sciences de la réadaptation, Université Laval, 1050 avenue de la Médecine, Québec City, QC G1V 0A6, Canada

[4] Observatoire de linguistique Sens-Texte, Université de Montréal, Montréal, QC, Canada

Published online: 07 January 2025

muv/" has a 100% probability of being followed by "/mɑ̃/", forming the word "mouvement" [movement]), then hearing "/muv/" provides valuable insights for anticipating "/mɑ̃/". Compare this to "/ma/", which only has a 6% probability of being followed by "/mɑ̃/", forming the word "maman" [mommy], an otherwise very frequent word in the French language. The difference is that /ma/ is a more frequent and more productive syllable than /muv/ in French, which makes it much harder to predict the syllable that follows it. This predictive capability can aid in clarifying speech in challenging listening conditions (for instance in noisy environments) or when listening to talkers with unfamiliar accents.

This initial work thus opened the way for new investigations of lexical (words, lemmas) and sublexical (morphological) phenomena that can inform a variety of studies in various fields (psycholinguistics, education, psychology, linguistics). There was indeed a paucity of databases providing full corpus analysis of language at all aforementioned levels in French, and especially in Québec French, more particularly in the spoken register. Frequency values of written and spoken Québec French words were published 30 years ago (Baudot, 1992; Beauchemin et al., 1992; Séguin, 1993), but these reports did not include analyses of sublexical properties. The *Phonologie du français contemporain* database (Durand et al., 2002) includes data from speakers of several French-speaking countries but only a small number of Québec French speakers ($n=31$). Furthermore, this database does not include tools to calculate statistics about lexical and sublexical units (e.g., syllables, morphemes, words).

The alternative for researchers working on Québec French is to use data available for French spoken in France. Considering that there are important differences between the two varieties (Poirier, 2009; see also https://www.tlfq.org), this alternative has several limitations. The best-known database is LEXIQUE 3 (New, 2006; New et al., 2001). LEXIQUE 3's spoken language corpus consists of film subtitles and therefore does not represent spontaneous spoken language. LEXIQUE 3 is fully lemmatized, and it is possible to obtain frequency values for both lemmas and tokens. However, LEXIQUE 3 provides information for a subset of derivational morphemes only (Namer, 2003a, b). Smaller corpora have been fully analyzed for morphology, but frequency statistics from those sources are influenced by the corpus' topic and lexical field (Fradin et al., 2008). Several databases of morphological variables from corpora of French spoken in France have been made available in the last 15 to 20 years (see Mailhot et al., 2020, for a description of MorphoLex and review of other databases based on diverse corpora).

The databases described above report linguistic analyses of morphology, but none report data about the actual perception of morphological phenomena by contemporary speakers of French, such as semantic transparency. Semantic transparency refers to the capacity to determine the meaning of a word based on its constituents (Marslen-Wilson et al., 1994). It is often studied by contrasting the processing of morphologically complex words ("teacher"—"teach") with the processing of pseudo-morphological words ("corner"—"corn") (Diependaele et al., 2012; Jared et al., 2017). Another aspect of semantic transparency concerns the possibility to identify the parts that form complex words (i.e., morphemes) and access their meaning. In the present article, we report perceived semantic relatedness between morphologically complex words and their roots.

To facilitate research on spoken language across a variety of disciplines (e.g., psycholinguistics, experimental phonetics, cognitive neuroscience of language, phonology, corpus linguistics, experimental psychology), we also updated our corpus of contemporary spoken Québec French, *SyllabO+* by adding an additional 41 speakers (for a total of 225 speakers) and creating three new databases (unique words, lemmas, and morphemes), which we describe in this article. The morphological analysis included inflectional, derivational, and compositional morphology. We also conducted an online study to assess the semantic transparency of derived words that are part of SyllabO+. As a result, SyllabO+ now includes six databases: phones, syllables, words (new), lemmas (new), morphemes (new), and an index of semantic transparency (new).

In this article, we describe the corpus updating process, the four new databases, and the semantic transparency study, as well as potential use of these tools in various fields including teaching, first and second language learning, speech-language pathology, and language research.

Methods

Corpus update

The initial corpus, published in 2017, contained 184 adult speakers recorded from 2012 to 2016, to which we added 41 speakers later in 2017, leading to a corpus of 225 speakers. Out of the 41 new speakers, 22 were recorded in formal context, while 19 were recorded in informal contexts. Formal contexts included interviews, lectures, press conferences, and radio/television programs, primarily sourced online from public media archives (2000–2016). Although speakers in formal settings tend to use more "standard" speech, we selected samples that were spontaneous rather than read aloud. Additionally, 7% of formal samples were recorded by our team in lectures, lab meetings, or conferences. Most informal recordings were conducted by our team (in the lab or at participants' homes) between 2013 and 2016, with 3% sourced from online resources. During lab recordings,

Table 1 Number of syllables and words transcribed by age, sex, and communication context (formal, informal)

		18–45 years		46–70 years		71–100 years		Total
		Male	Female	Male	Female	Male	Female	
Formal	No. speakers	28	25	26	26	18	15	138
	Syllables	29,405	29,805	30,865	34,216	32,232	18,591	175,114
	Words	23,353	22,523	23,529	26,195	25,536	15,485	136,621
Informal	No. speakers	12	12	11	11	19	22	87
	Syllables	27,126	28,744	27,324	28,643	36,396	40,955	189,188
	Words	24,721	27,365	24,102	26,096	30,993	35,707	168,984
Total	Syllables	56,531	58,549	58,189	62,859	68,628	59,546	364,302
	Words	48,074	49,888	47,631	52,291	56,529	51,192	305,605

a Lavalier microphone was used, and participants were encouraged to discuss self-selected topics, allowing the conversation to flow naturally.

The final corpus includes samples from 225 different speakers (representing 364,302 syllables and 305,605 words; refer to Table 1 for the details), including 114 male and 111 female speakers, recorded in either formal (61%) or informal (39%) communication contexts. Formal samples represent 61% of the samples in the corpus, 48% of the total syllable count, and 45% of the total words in the corpus. Informal samples represent 39% of the samples in the corpus, 52% of the total syllables, and 55% of the total words.

The speakers were native speakers of Québec French[1] (54.5 ± 19.6 years, range 20–97 years), with a mean of 16.5 ± 3.9 years of education[2] (range 7–27 years). The participants were divided into three groups: 18–45 years (mean 32 ± 6.8 years; N = 77), 46–70 years (mean 55 ± 7.6 years; N = 74), and 71–97 years (mean 78 ± 6.4 years; N = 74). Participants were recruited through written ads posted in the community (e.g., supermarkets, coffee shops, drugstores, hospitals, and websites), emails to large groups (e.g., university students and staff, senior groups), presentations in retirement centers, and by contacting people on the Lab participant database.

The study was approved by the *Comité d'éthique de la recherche sectoriel en neurosciences et santé mentale, Institut Universitaire en Santé Mentale de Québec* (#356–2014). The XML transcription and all databases are freely available on the SyllabO website (https://syllabo.speechneurolab.ca) as well as on Borealis, the Canadian Dataverse Repository, at https://doi.org/10.5683/SP3/T3ZUIN. The original voice recordings cannot be shared because at the time the recordings were made, participants did not consent to public data sharing of their voice.

Corpus transcription and syllabification

These steps are identical to those detailed in our previous article and will not be described in the present article, which focuses on the creation of new databases (Bédard et al., 2017). The recordings were transcribed both orthographically and phonetically using International Phonetic Alphabet (IPA) symbols. We did not transcribe prosodic features, silences, laughs, non-linguistic onomatopoeia, or background noise (non-speech elements). All transcription protocols (orthographic, phonetic and syllabification) are provided on our website (www.speechneurolab.ca/syllabo), under "Documentation". The phonetic transcriptions were next syllabified, that is, split into syllables. As illustrated in Table 1, the number of spoken syllables in the phonetic transcription was smaller than the number of orthographic syllables. This is because many sounds are elided in spoken French, resulting in fewer syllables. For example, "je suis" [I am] is often pronounced /ʃy/ or /ʃyi/ in spoken French; hence, two words are produced as one spoken syllable. Likewise, "je l'avais" [I had it] can be pronounced /ʒlavɛ/, resulting in two spoken syllables for a three-syllable orthographic syntagm. Even words that are considered as three syllables in a dictionary can become two syllables when spoken (e.g. "médecin" [physician] is usually pronounced as /metsɛ̃/).

Creation of the Word and Lemma databases

The syllabified and the orthographic transcriptions were saved as annotated and marked-up XML files. All metadata were anonymized and saved in a separate XML file that was linked to each individual transcription by a reference number. Extracting statistical information from these XML files was done by means of a Python script. The extracted

[1] Speakers were born in Québec and reported Québec French as their native language (language learned at home via parents speaking Québec French).

[2] The number of years of education was calculated only with speakers for which we had these data, i.e., participants recorded by our team (N = 106).

statistical information was organized in tables, which constitute the databases. The Word and Lemma databases were integrated into the SyllabO+ web application in 2018.

Next, we used a Python script and the open-source TreeTagger tool[3] to carry out two processes: tokenization and lemmatization. Tokenization is the process of converting text into individual words or tokens, while lemmatization, is the process of converting words to their base or root forms also called "lemma" (canonical form). Each word was included in the Word database, along with its grammatical gender, number, and conjugation marks. Uppercase letters were converted into lowercase letters. The output of TreeTagger was inspected and manual interventions were made whenever necessary. Following tokenization, lemmatization was conducted. For example, the French sentence "On capture aussi les poissons au filet et à la ligne" [translation: Fish are also caught with nets and lines] becomes: "on / capturer / aussi / le / poisson / au / filet / et / à / le / ligne" [fish / be / also / catch / with / net / and / line].

Again, the output of TreeTagger was inspected and post-processing interventions were made whenever necessary. The cases that were flagged by TreeTagger as ambiguous were verified and the appropriate lemma was selected. For the cases that were flagged by TreeTagger as <unknown>, the original word was used.

The Word and Lemma databases each consist of three different data tables: unique words/lemmas with related data and statistics, and word/lemma collocations, which include pairs of words/lemmas with related data and statistics and groups of three words/lemmas with related data and statistics. A description of the database tables, with definitions and detailed description of calculations, is available on our http://syllabo.speechneurolab.ca, under "Documentation". The complete databases, or a specific subset of the databases resulting from specific query options, can be downloaded from the web application. The following parameters can be used individually or in combination: context of communication (formal, informal), age (range), and sex of the speakers. Files are downloaded in CSV format (Comma Separated Values), which is a way of storing tabular data in plain text—in this case, UTF-8 text.[4]

Morphology

The unique word database was the starting point for investigating inflectional morphology, derivational morphology, and composition. One person (AL) oversaw all word segmentation analysis and word coding in the database. Each word segmentation was verified by the first author (NAO) to finalize the analysis and prepare the morphological transparency test.

Inflectional morphology

Inflectional morphology concerns the addition of inflectional markers to roots to signal morphosyntactic information and relationships between words (Lehmann & Martin-Berthet, 2005). The types of morphosyntactic information that can be specified for words depend on their grammatical category. In French, verbs can carry information about the person (first, second, third), number (singular, plural), tense (present, past, future), gender (feminine, masculine—for the past participle only), mood (indicative, subjunctive, imperative, conditional,[5] infinitive, participle), and aspect (perfective, imperfective[6]). Pronouns can carry information about the person[7] (first, second, third), gender (masculine, feminine), and number (singular, plural). Determiners, nouns, and adjectives can carry information about gender (masculine, feminine) and number (singular, plural). Prepositions, conjunctions, and adverbs do not vary grammatically and are not inflected.

The information generated by the analysis in TreeTagger was used as a starting point for the analysis of inflectional morphology and expanded with additional categories, as needed. For verbs, TreeTagger terminology was kept, but fields were added to distinguish the tense and the mood, and to add information about the person, number, and gender. The terminology is consistent with the one used in LEXIQUE3 (New, 2006; New et al., 2001). Table 2 provides examples of inflectional morphology analysis for a representative word from each of the grammatical categories.

Derivational morphology

Derivational morphology includes several word formation mechanisms, including transposition, back-formation, and affixation (Bauer, 2008; Lehmann & Martin-Berthet, 2005). Transposition (also called conversion or zero-derivation) consists of changing a word's grammatical category, without

[3] See http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/ for documentation.

[4] UTF-8 is the most common standard international encoding system to display all characters correctly, including accents or special characters.

[5] The conditional is sometimes considered a tense of the indicative mood.

[6] In French, aspect is only expressed in the past tense. The "imparfait" (imperfective) is contrasted with the "passé simple" and "participe passé" (perfective). Information about perfectiveness was aggregated with information about tense by using separate codes for those three inflections in the tense field.

[7] French pronouns show remnants of case marking (e.g., third-person masculine singular: "il" (subject), "le" (direct object), "lui" (indirect object)). However, they are considered separate words.

Table 2 Example of inflectional morphology analysis for one representative word from each grammatical category

Grammatical category	Word (lexeme)	Lemma	Mood	Tense	Person	Number	Gender
Verb	allaient	VER:impf,aller	ind	impf	3	p	N/A
Noun	bulles	NOM,bulle	N/A	N/A	N/A	p	f
Adjective	délicat	ADJ,délicat	N/A	N/A	N/A	s	m
Pronoun	auxquels	PRO:REL,auquel	N/A	N/A	N/A	p	m
Determiner	une	DET:ART,un	N/A	N/A	N/A	s	f
Preposition	sans	PRP,sans	N/A	N/A	N/A	N/A	N/A
Conjunction	car	KON,car	N/A	N/A	N/A	N/A	N/A
Adverb	plutôt	ADV,plutôt	N/A	N/A	N/A	N/A	N/A

SyllabO+ abbreviations: *f*, féminin (feminine), *impf*, imparfait (imperfective), *ind*, indicatif (indicative), *m*, masculin (masculine), *p*, pluriel (plural), *s*, singulier (singular)

TreeTagger abbreviations: *ADJ*, adjective, *ADV*, adverb, *DET:ART*, determiner—article, *KON*, conjunction, *NOM*, noun, *PRO:REL*, relative pronoun, *PRP*, preposition, *VER*, verb

transforming it. This word formation mechanism is common for verbs in English. It is also found in French (e.g., "savoir", verb (to know)—"savoir", noun (knowledge); "méchant", adjective (mean)—"méchant", noun (villain)). Back-formation consists of creating a new word by removing real or supposed affixes from its root. For example, in English, the verb "resurrect" was formed based on "resurrection". The existence of back-formation in French is more controversial. Words like "refus" (refusal) (from "refuser" (to decline)) and "cri" (a cry) (from "crier" (to shout)) have been considered as examples of back-formation in French. However, roots (or part of words identified as roots) rarely form words in French. For example, the verb "somnoler" (to doze off) is analyzed by some as being formed by back-formation from "somnolence", but to form "somnoler", the ending "-er" has been added to the root extracted from "somnolence". This ending can be seen as derivational (approximating this word formation to affixation, see below) or as inflectional ("-er" being the desinence of the infinitive). Some argue that words that are considered as formed by back-formation are actually formed by transposition from one of their inflected forms (Lehmann & Martin-Berthet, 2005). The third form of derivation, affixation, consists of adding one or more suffixes and/or prefixes to a root to create a new word. Affixes attach to roots from specific grammatical categories and can be more or less productive, that is, more or less likely to be used to coin new words. Affixes generally induce a predictable change in meaning to the root (e.g., verb+"-er" means *a person who* [verb]: bake-baker, swim-swimmer, teach-teacher, etc.) but they may have more than one meaning (e.g., verb+"-er" can also mean *an object that* [verb]: light-lighter, stick-sticker, dry-dryer, etc.), and they can be identical to initial or final syllables that are not morphemes (e.g., "-er" in "corner", "brother", "super", etc.).

The analysis of derivational morphology was performed according to the comparative lexical analysis of French words reported in Le Robert brio (Rey-Debove, 2004). Le Robert brio provides a morphological analysis of 33,000 words, lists of suffixes and prefixes, and definitions of affixes with reference to words in which they are included. After excluding one-syllable words and function words, all words in the SyllabO+ corpus were analyzed for morphological structure. Words were identified as derived or not, and affixes and roots were identified and segmented to provide detailed information about internal word structure.

Composition

Composition (or compounding) is a word formation mechanism that consists of joining two or more existing words to create a new word (e.g., "toothbrush", "takeout", "blackboard", etc.). It is generally considered more common and more productive in Germanic languages, such as English, than Romance languages, such as French (Arnaud & Renner, 2014). In French, words forming a compound can be fused (e.g., "portefeuille" (wallet)), connected with a dash (e.g., "lave-vaisselle", (dishwasher)), or separated by spaces (e.g., "rouge à lèvres" (lipstick)). The status of compounds is debated in French. Some linguists consider that only word combinations that cannot be formed using syntactic rules can be considered as compounds (Fradin, 2011).[8] Others include word combinations that form cohesive lexical units (Abeillé & Clément, 2003; Mathieu-Colas, 1996), supported for example by showing that inserting adjectives or adverbs within a potential compound creates infelicitous syntagms (e.g., "*[rouge <u>rose</u> à lèvres]" vs. "[rouge à lèvres] <u>rose</u>"

[8] Fradin acknowledges that a sudden increase in lexical co-occurrence (e.g., "guerre froide" (cold war) in the 1950s) might indicate a cohesive lexical status despite the possibility to form the unit using syntactic rules.

(pink lipstick)). In this study, we adopted the more inclusive definition of compounds that includes all word combinations forming cohesive lexical units.

Words were classified as compounds (“lave-vaisselle”) or non-compound words (“échelle”). Compounds formed with words connected by a dash were analyzed based on their orthographic transcription. Fused compounds were analyzed based on the information available in Le Robert brio (Rey-Debove, 2004). Exploratory co-occurrence analyses were conducted on the original corpus transcripts using the TermoStat tool (Drouin, 2003) to extract compounds separated by spaces. Some extracted examples include “ours polaire” (polar bear), “cuir chevelu” (scalp), and “chou de Bruxelles” (Brussels sprout). However, a large proportion of examples consisted of syntagms of various lengths that did not meet other criteria to be considered as compounds (e.g., the insertion test described above). To be fully informative, this analysis would likely have required a larger corpus and a systematic comparison of information extracted from other large corpora, which went beyond the scope of this study.

Other word formation mechanisms and special cases

Numbers Numbers were noted as compounds when appropriate, but their internal structure was not further analyzed, and the code NUM was entered in the internal structure column.

Proper names Proper names were noted as compounds when appropriate (e.g., “Chaudière-Appalaches”, the name of a region in the province of Québec), but their internal structure was not further analyzed, and the code NAM was entered in the internal structure column.

Acronyms and words derived from acronyms Acronyms were counted as compounds. However, their internal structure was not further analyzed, and the code ACR was entered in the internal structure column. Words that are derived based on an acronym, such as “péquiste” (a member or supporter of the Parti Québécois, PQ) were segmented based on the acronym and affix (péqu (pq) – iste).

Words from other languages Words from other languages were not analyzed for their morphological structure. This includes derived words (e.g., “castillano”, “computer”), compounds (e.g., “talk-show”, “weekend”), and borrowings formed with roots from a different language and productive French affixes (e.g., “bruncher” /bɹʌnʃe/ (to brunch), “peopleisation” /pipəlizasjɔ̃/ (in the media, reporting on the private life of individuals who are not in entertainment—especially politicians—as if they were celebrities). The code ETRAN was entered in the grammatical category column for those words.

Words specific to Québec French Words that are not used in other varieties of French (e.g., “tripper”, to have a good time) and for which etymology was uncertain were not analyzed. The code QUEB was entered in the grammatical category column.

Portmanteaus (blends) Words like “courriel” (email), from “courrier” (mail) and “électronique” (electronic), were segmented based on the morphological structure of their component words (courri (courrier)—el (électronique)).

Clippings (truncations) Words like “coloc” (roommate), from “colocataire”, were segmented based on the structure of their corresponding complete word (co – loc (locataire)).

Onomatopoeia The code ONOM was entered in the grammatical category column.

Errors Speakers occasionally produced errors (e.g., *“acquéri” instead of “acquis” for the past participle of the verb “acquérir” (acquire)). Errors were labeled “Errors” and were not analyzed.

Semantic transparency study

In addition to providing a linguistic and etymological analysis of morphologically complex words, the present study aimed to provide an estimate of the semantic transparency of derived words from speakers of Québec French. Semantic transparency refers to the semantic relatedness of a morphologically complex word and its root (Marslen-Wilson et al., 1994; Marslen-Wilson & Zhou, 1999). While analyzing a word’s history (that is, its etymology) allows linguists to know that words have been formed by derivation, those words are not always perceived as such by contemporary speakers (Rey-Debove, 2004). Derived words can take on a meaning of their own, so much so that people no longer relate their meaning to that of their root. If a word’s morphological structure goes unnoticed, speakers cannot take advantage of the support provided by the meaning of the root to understand the full word or deduct the meaning of the full word based on the meaning shared by other words formed with the same prefix or suffix. This has several implications for word processing, since morphology plays an important role in improving reading fluency and comprehension in people who have challenges decoding written words (i.e., dyslexia) (see for example Casalis et al., 2004; Cavalli et al., 2017; Elbro & Arnbak, 1996; Martin et al., 2013), and in improving vocabulary and other language skills in one’s native (Ashkenazi et al., 2020; Bowers & Kirby, 2010; Carlisle et al., 2010; Singson et al., 2000) and second language (Brooks et al., 2011; Kimppa et al., 2019; Lam & Chen, 2018).

Speakers of Québec French provided semantic transparency judgments for a subset of words derived from SyllabO+. Since we only used morphologically complex words, the objective aspect of this judgment was controlled: all pairs of derived words and roots included in the study could objectively be rated as related, based on their linguistic and etymological analysis. We were interested in participants' subjective perception of these word pairs. The selection of derived words and methods for this study are described below.

Selection of words for transparency judgments

To obtain estimates of semantic transparency, speakers were asked to judge to what extent a derived word and its root are related in meaning. In order for this judgment to rely on semantics and not on metalinguistic awareness, roots need to be real words that function independently in the language. For example, the pair "minuit" (midnight) and "nuit" (night) lends itself well to semantic transparency judgments, while "midi" (midday, noon) and *"di" does not, because *"di" is not a word. It is found in "diurne" (diurnal), "quotidien" (daily), and other words with a meaning related to "day" as a vestige of the Latin origin of French, but the French word for "day" is "jour".[9] All derived words were screened to identify suitable roots for semantic transparency judgments.

In words with multiple affixes, the shortest unit was used as the root (e.g., for "re-lâche-ment" (release, relapse; noun) the root was "lâcher" (let go) and not "relâcher" (release, relapse; verb)). Contrary to English, most French words have affixes. The infinitive was used for verb roots (e.g., "relâchement" was paired with "lâcher"). Words that underwent phonological changes through affixation were included in the study (e.g., "méchant" /meʃɑ̃/ (mean), "méchanceté" / meʃɑ̃ste/ (meanness), "méchamment" /meʃamɑ̃/ (meanly)) (Marslen-Wilson et al., 1994; Marslen-Wilson & Zhou, 1999). In order to maximize the information available in the database, French words were used instead of Greek and Latin roots when possible (e.g., "memor—" / "mémoire"). To be used instead of the root, the French word needed to be semantically close to the root, which was shown for example by being used in the definition of the root in the Le Robert brio (Rey-Debove, 2004). To ensure that judgments were based on morphological relationships and not on lexical associations, the French word and the Greek or Latin root also needed to have sufficient phonological overlap. To measure phonological overlap, Greek and Latin roots and their corresponding French word were transcribed phonologically. Phonemes were divided into three categories: those found in both the root and the French word, those found only in the root, and those found only in the French word. The total number of phonemes found only in the root or the French word was subtracted from the total of overlapping phonemes. Systematic or quasi-systematic patterns between languages such as French nasal vowels corresponding to vowels with the phoneme /n/ in Latin roots were not treated separately and were considered to contribute to the overall phonological difference between the root and the word. Pairs with a score of at least 1 were kept to use the French word in semantic transparency judgments (e.g., /mɛmɔr/, /mɛmwar/ - eight phonemes in common, one in the Latin root only (/ɔ/), two in the French word only (/w/, /a/): six in total > 1). Other examples of words that were kept for the test include "bref" for "briev—", "fête" for "fest—", and "règle" for "regul—". Examples of words that were rejected include "nuage" for "nebul—", "air" for "aero-", and "égal" for "equi-". In total, 3764 words were included in the study of semantic transparency of derived words.

Contrary to other studies (Jared et al., 2017), pairs of synonyms and pairs composed of a pseudo-root and a pseudo-derived word were not included in the transparency judgment task. The main goal of the present study was to provide an estimate of semantic transparency for a large number of derived words. Because there was already a large number of words to include in the study, we did not want to include fillers that would unduly extend task completion time. Furthermore, judgments made on synonyms and pseudo-derived words likely compress the range of scores given for true derived words. Even if it is semantically transparent, a pair composed of a derived word and its root will not be considered as semantically related as a pair of synonyms. Moreover, rating the semantic relationship between pseudo-roots and pseudo-derived words tests a different kind of knowledge than rating the relationship between real roots and real derived words. Judgments of pseudo-derived words may implicitly rely on knowledge of which affixes go with which base, and of the semantic impact of affixes on the root. The goal of studies that test the judgment of pseudo-morphological words and synonyms has more to do with morphological awareness and have different goals from those of the present study.

Compounds were not included in the study of semantic transparency. As mentioned above, the definition of compounds is controversial in French (Fradin, 2011). Furthermore, semantic transparency judgments for compounds depend on the relationship between the compound and its head, which is different from and cannot be equated to the relationship between a derived word and its root.

[9] Other similar examples include "fat(i)-" in, for example, "fatidique" (fateful) (cf., the English word "fate"; "destin" in French) and "pon(ent)-" in, for example, "disponible" (available) (cf., the Spanish word "poner" (put); "mettre" in French).

Participants

The study was completed online using LimeSurvey (https://www.limesurvey.org/). Participants were recruited by posting messages on social media and on personal and research center websites, and by sending messages through mailing lists. Participants provided their consent to participate by clicking on a link that led them to the survey. Participants needed to be 18 or older and to have sufficient knowledge of French to participate. The study was approved by the Institutional Review Board of the Faculty of Medicine and Health Sciences at McGill University (#A04-E24-20B) and the *Comité d'éthique de la recherche sectoriel en neurosciences et santé mentale, Institut Universitaire en Santé Mentale de Québec* (#2021–2143). It was conducted according to the principles of the Declaration of Helsinki, and all participants provided informed consent to participate.

A total of 490 participants completed the survey. Forty-five participants had a native language other than French, and their data were analyzed separately. Four participants only completed the sociodemographic portion of the survey; their data were eliminated from further analysis. The final sample thus included 441 participants. Participant characteristics are reported in Table 3.

Table 3 Sociodemographic characteristics

Age (years)	
Mean (SD)	36 (15.0)
Sex	
Male	21.30%
Female	77.60%
Nonbinary, other	0.90%
Education	
Elementary school	0%
High school	1.80%
Vocational degree	2.50%
College/CEGEP	27.70%
Bachelor's degree	31.50%
Master's degree	24.30%
Doctorate (MD, PhD)	11.10%
Learning history	
Consulted a specialist	13.40%
Did not consult a specialist	86.60%
Country of birth	
Canada	83.70%
France	12%
Other	4.30%
Bilingualism/multilingualism	
Speaks French only	12.70%
Speaks French and other languages	86.80%
Number of languages spoken	
Mean (SD)	2.47 (0.89)

Material

The online survey was divided into two sections. The first section consisted of a sociodemographic questionnaire that included questions about personal characteristics and language background. Questions probed for age, sex, highest level of education completed, consultation with a learning specialist before the age of 12, country of birth, first spoken language, other spoken languages, and level of comprehension, expression, reading, and writing proficiency in each spoken language. The second section of the survey consisted of a semantic association judgment task. Briefly, participants were asked to rate whether they agreed that two words were related in meaning using a scale from 1 (completely disagree) to 7 (completely agree). Pairs of words consisted of a derived word and its root. Participants were asked to provide ratings on 130 word pairs. The full list of 3764 derived words was divided into 29 lists, with a list that included words that were repeated to ensure that all lists included 130 items. The distribution of words across the different lists was done controlling for frequency and affixation type. To do so, the full list was divided into 20 bins of frequency using the information collected in the lemma analysis stage (see Creation of the Word and Lemma databases). Words were also coded for word-initial and word-final letters. Then, a pseudo-random order of the full list constrained for repeats of words from a specific frequency bin, a specific initial letter, and a specific final letter was generated using Mix (van Casteren & Davis, 2006). Three versions of each list were created by generating pseudo-randomized orders of words.

Participants were randomly assigned to one of the 29 lists in one of its three versions after completing the sociodemographic section of the survey. Participants who reported having a native language other than French were assigned to a different list that included some of the most frequent derived words in SyllabO+, while maintaining diversity in terms of affixation.

Procedure

Participants completed the sociodemographic questionnaire and were asked to click on a button to continue. They were presented with the task instructions. Briefly, participants were told to judge whether two words were related in meaning. They saw a pair of words and the prompt "These words are related in meaning". Answer choices ranged from 1 (completely disagree) to 7 (completely agree). Participants made their choice by clicking on a box next to the choice. Participants were not told about the morphological relationship between words in each pair, but they were told to focus

Table 4 Summary statistics for semantic transparency of derived words

	Number of respondents	Semantic transparency					
		M	*SD*	Median	IQR	Min	Max
Mean	14.77	5.37	1.35	5.61	1.66	2.81	6.87
Standard deviation	4.09	1.36	0.59	1.67	1.19	1.82	0.56
Minimum	8	1.18	0	1	0	1	2
Maximum	42	7	2.74	7	6	7	7
First quartile	10	4.59	0.82	5	1	1	7
Median	15	5.80	1.41	6	1.5	2	7
Third quartile	17	6.44	1.85	7	2.25	5	7
IQR	7	1.85	1.03	2	1.25	4	0
20th percentile	10	4.19	0.72	4	1	1	7
40th percentile	13	5.44	1.18	6	1	2	7
60th percentile	16	6.12	1.59	6.5	2	3	7
80th percentile	17	6.53	1.93	7	2.75	5	7

M, mean; *SD*, standard deviation; *IQR*, interquartile range

on the semantic relationship between words and to ignore similarities in orthography and pronunciation.

Three examples were provided before the beginning of the task. Words in the example were not included in the main lists. The first example illustrated a transparent semantic relationship between a word and its root (“chat”—“chaton”; cat—kitten). Participants were told that their rating for this pair could be 6 or 7. The second example included words that are not related morphologically, but that are related phonologically and orthographically (“pas”—“repas”; step, not—meal). Although we did not want to include pseudo-morphological pairs in the main lists, we deemed it more appropriate to use a non-related pair in this example to avoid biases. Participants were told that possible responses for this pair could be 1 or 2. The third example used a polysemous word to illustrate that semantic ratings could vary depending on perception and that there were no inherently good or bad answers (“tour” — “contour”, border, turn, tower, trick, etc.—contour, outline, edge, rim, etc.). Participants were told that their rating could fall anywhere on the scale for this pair. We deemed it more appropriate than instructing participants to make sure to use every point on the scale when completing the survey. Once they were done, participants clicked on a button to transmit their results. If they wished, they were invited to enter their email address to participate in the drawing for a gift card.

Data analysis

Results were downloaded in .CSV format from the LimeSurvey website. Spreadsheets were organized and merged to collect all data pertaining to specific words from each version of the lists and each list in the survey. Data were compiled in long form, detailing all information for each word and each participant. Average summaries per word and per participant were also compiled.

Because of limitations in the list assignation algorithm used in LimeSurvey, more participants provided ratings for some lists than others. We compiled the number of people who provided ratings for each word. That number ranged from 8 to 42 (mean: 14.77; standard deviation: 4.09). We report this information in the database so users of SyllabO+ can use it to select items or control for this variable in their analyses.

Data were visualized using scatterplots and boxplots generated using RStudio. We calculated statistics to characterize central tendency, dispersion, and distribution of semantic transparency ratings. Because there is no consensus on the categorial or continuous nature of measures using a Likert response format (Carifio & Perla, 2007; Jamieson, 2004), we report statistics consistent with the categorial and the continuous view. Generalized linear mixed models, analysis of variance, and regressions were used to analyze the distribution of transparency ratings, the relationship between sociodemographic characteristics (age, sex, education, learning history, and number of languages spoken) and French proficiency and transparency ratings, and the relationship between the type of affixation (prefixation, suffixation) and transparency ratings. Analyses were conducted in R version 4.0.2. and SPSS version 23.

Results

Results for the semantic transparency ratings are provided below. Since these results can be used by readers to inform original studies using words included in the SyllabO+

database, we report results using a perspective centered on the word (rather than the participant) as the unit of observation.

Transparency ratings

Transparency ratings statistics are reported in Table 4. The table shows summary statistics for values compiled for the 3764 words included in the analysis.

The majority of words are of medium to high transparency, but statistics reveal variability in ratings. SyllabO+ users can use those statistics to guide the selection of words for a variety of experiments. However, users are encouraged to carefully consider variability for different levels of transparency, as will be explained below.

Distribution of transparency ratings

Standard deviations and interquartile ranges (IQR) reported in Table 4 show variability in semantic transparency ratings. To check whether variability was similar at all levels of transparency, we plotted standard deviations according to mean transparency using a scatter plot (Fig. 1) and IQR according to median transparency using violin plots, which provide a full understanding of data distribution (Fig. 2). Both figures show that variability is greater for words of medium transparency. Figure 1 illustrates the quadratic relationship between mean semantic transparency and standard deviation. A quadratic equation explains 82% of the variance, $F(2, 3.76) = 8682.01$, $p < .001$.

Transparency ratings and affixation type

Past research has shown that suffixed words are more transparent than prefixed words. We coded words in the study as prefixed, suffixed, and both prefixed and suffixed. For this analysis, we considered that words that could be analyzed as being formed by back-formation or by a transposition of an inflected form were suffixed or both prefixed and suffixed, when relevant (e.g., “voler” → “survol”: both). We ran a one-factor ANOVA of mean transparency by affixation type using Dunnett’s test for post hoc comparisons as an alternative to the Bonferroni test because homogeneity of variance was not assumed. The overall difference between affixation types was significant, $F(2) = 480.844$, $p < .001$. Suffixed words were more transparent than prefixed words and words that are both prefixed and suffixed. Prefixed words were also more transparent than words that are both prefixed and suffixed (all post hoc comparisons: $p < 0.05$). We repeated the same analysis with mean transparency and obtained similar results. Figure 3 shows mean transparency by affixation type in the form of violin plots.

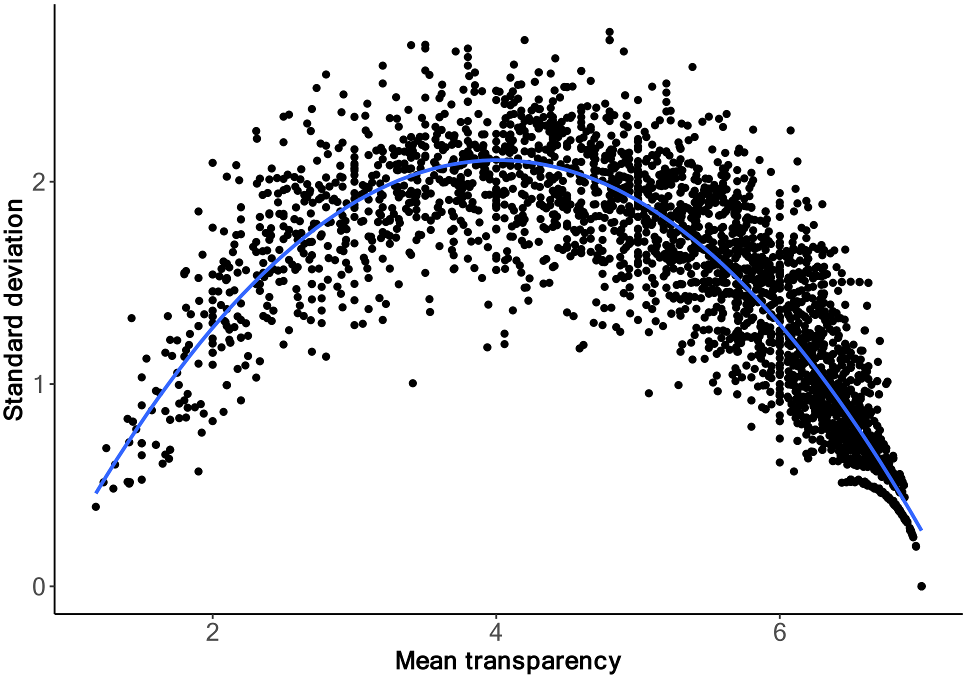

Fig. 1 Mean transparency and standard deviation for the 3764 words included in the analysis

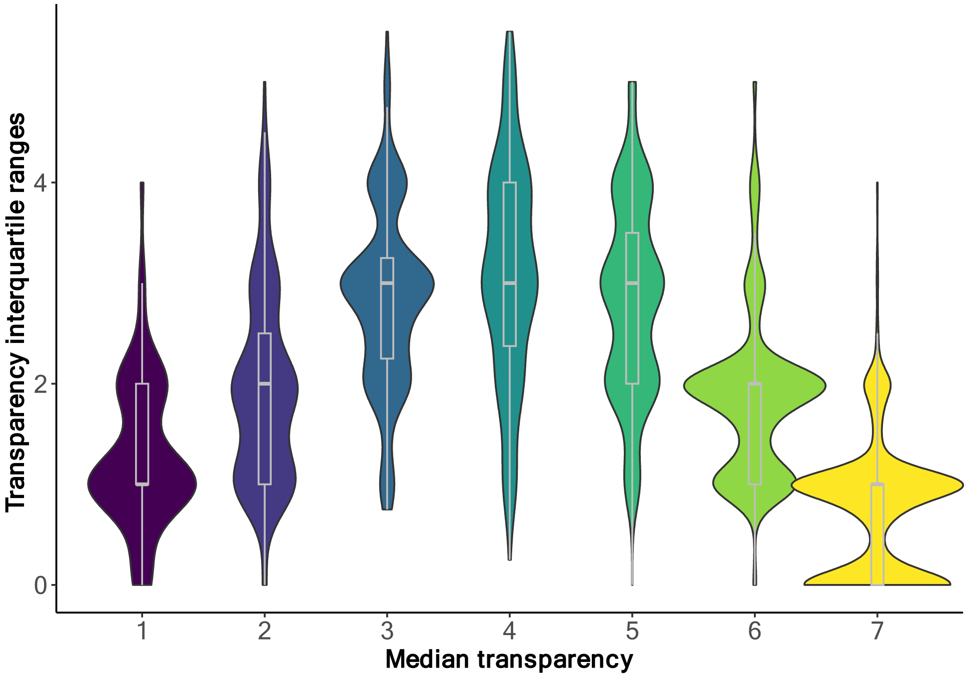

Fig. 2 Violin plots showing the distribution of median transparency and interquartile range (IQR) for the 3764 words included in the analysis

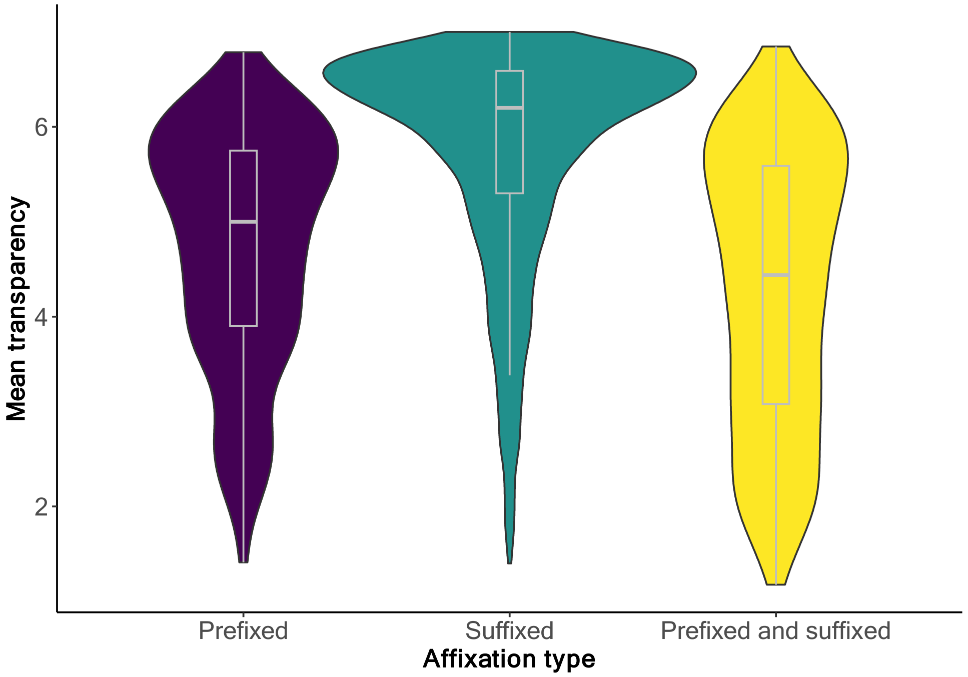

Fig. 3 Violin plots showing the distribution of word transparency as a function of affixation type

Transparency ratings and sociodemographic characteristics

To study the impact of sociodemographic characteristics on transparency ratings, we fitted a generalized linear mixed model of transparency ratings with age, sex, education, learning history, and number of languages spoken as fixed factors, and random intercepts and slopes for participants and items. The model was fit by maximum likelihood. Significance was tested with *t*-tests using Satterthwaite's method. An exploratory correlation analysis showed a weak but significant correlation between age and education, and age and learning history, which were also correlated to average semantic transparency. Therefore, we modeled an interaction between learning history, age, and education, but not those factors and sex. Using alpha = .05 as the significance threshold, we found no significant main effect or interaction. The effect of education was $p = .069$. Details are reported in Table 5.

Transparency ratings in speakers of French as a second language

Forty-five participants reported speaking French as a second language. They collectively spoke 16 different native languages grouped in eight language families—Atlantic-Congo languages: Ngiemboon ($n = 1$), Wolof ($n = 1$); Berber/Amazigh languages: Amazigh ($n = 1$), Kabyle ($n = 1$); Chinese languages: Cantonese ($n = 2$), Teochew ($n = 1$); Germanic languages: Dutch ($n = 1$), English ($n = 9$), German ($n = 3$); Indo-Aryan languages: Bengali ($n = 1$); Iranian languages: Persian ($n = 2$); Romance languages: French-based Haitian Creole ($n = 1$), Portuguese and Brazilian Portuguese ($n = 5$), Romanian ($n = 1$), Spanish ($n = 9$); Semitic languages: Arabic ($n = 6$). They were aged between 19 and 80 years ($M = 33.8$; $SD = 12.1$), were 68.89% female, and were highly educated, with 80% having completed at least an undergraduate university degree. They spoke between two and four languages ($M = 3.4$; $SD = 0.62$). Their self-reported level of French proficiency (on a scale of 1 to 7, 7 being native-like) varied but was high overall—oral comprehension: 6.09 (1.22); oral expression: 5.91 (1.36); reading: 6.11 (1.32); writing: 5.81 (1.3).

Participants who had French as a second language provided ratings for 130 of the most frequent words included in the survey. Their ratings were compared to those of native French speakers for the same 130 words. There was no significant difference in average group ratings ($p > 0.05$). The influence of individual levels of French proficiency on the similarity of semantic transparency ratings was also tested. To do so, we computed the absolute difference between the rating of French as a second language participants and the rating of native French speakers for the same words. We fitted a generalized linear mixed model of absolute difference in transparency ratings with French oral comprehension proficiency, French oral expression proficiency, French reading proficiency, and French writing proficiency as fixed factors, and random intercepts and slopes for participants and items. Having established that sociodemographic variables did not have an influence on semantic transparency ratings, we did not include those factors in the model. The model was fit by maximum likelihood. Significance was tested with t-tests using Satterthwaite's method. Using alpha = 0.05 as the threshold, we found effects of reading and oral comprehension proficiency, but not oral expression and writing proficiency. Details are reported in Table 6.

Table 6 Effect of French proficiency on the absolute difference in semantic transparency ratings of speakers of French as a second language

	b	*SD*	*df*	*t* value	*p*
(Intercept)	1.57	0.21	45.58	7.57	< 0.001
French reading	−0.31	0.10	42.35	−3.12	0.003
French oral comprehension	0.24	0.12	42.30	2.03	0.049
French writing	0.06	0.08	42.52	0.69	0.49
French oral expression	−0.03	0.09	42.29	−0.36	0.72

b, regression coefficients; *SD*, standard deviation; *df*, degrees of freedom

Table 5 Effect of sociodemographic characteristics on semantic transparency ratings

	b	*SD*	*df*	*t* value	*p*
(Intercept)	4.44	0.51	257.88	8.71	< 0.001
Sex	−0.13	0.09	422.87	−1.53	0.13
Age	0.01	0.01	172.74	1.19	0.23
Education	0.19	0.10	276.30	1.83	0.07
Learning history	0.87	1.94	399.48	0.45	0.66
Age × Education	0.00	0.00	168.41	−0.91	0.37
Age × Learning history	−0.03	0.06	339.81	−0.53	0.60
Education × Learning history	−0.20	0.39	398.82	−0.52	0.60
Age × Education × Learning history	0.01	0.01	318.30	0.44	0.66

Sex (male, female, other); Age (years); Education (elementary school, high school, vocational degree, college-CEGEP, bachelor's degree, master's degree, doctoral degree); Learning history (consulted a learning specialist before the age of 12—Yes, No)

b, regression coefficients; *SD*, standard deviation

Discussion

In this study, the SyllabO+ corpus was expanded by adding 41 speakers and creating three new databases (unique words, lemmas, and morphemes). We also conducted a study with over 400 speakers of Québec French to assess the semantic transparency of 3764 derived words from the corpus. The addition of speakers and databases represents a substantial expansion of this tool and makes it an invaluable resource to inform an even larger range of psycholinguistic studies. One major strength of SyllabO+ is its focus on spontaneous oral language, which makes it representative of common language use. The study of the semantic transparency of derived words complements the linguistic analysis of morphology by providing information on the subjective perception of words' morphological structure by contemporary speakers of Québec French. Results from this study have several implications for studies of derivational morphology.

The summary of transparency statistics showed that ratings of transparency were relatively high overall. This was expected, since the lists only included word pairs that had a morphological relationship based on a systematic linguistic analysis of internal word structure. It is perhaps surprising that a non-negligible proportion of words (close to 20%) had ratings *below* the midpoint of the rating scale. This emphasizes how linguistic analysis can differ from the perception of contemporary speakers. The meaning of derived words evolves, sometimes independently from that of their root. Pairs of derived words that are redundant from an etymological and morphological point of view but distinct from a lexical/semantic point of view also illustrate this point (e.g., consider the pairs "nuageux" (cloudy) and "nébuleux" (nebulous), and "poussiéreux" (dusty) and "poudreux" (powdery)).

The summary statistics and inspection of semantic transparency ratings revealed considerable inter-individual variability. Variability was greatest for words that were rated as having medium transparency (4/7). The analysis of distributions showed that words that received a mean or median transparency rating close to 4 were not judged to be of medium transparency by most respondents. Instead, they were judged as very transparent by some, and as not transparent at all by others. Researchers using psycholinguistic tasks with derived words should not assume that all speakers of a language perceive the same degree of semantic transparency (Medeiros & Dunabeitia, 2016). Using the data reported in the present study, it is possible to select words whose ratings are consistent. Researchers can also choose to ask participants to provide semantic transparency ratings for the derived words used in their experiments and compare them to those reported in SyllabO+. However, it is important to acknowledge as a limitation that the variability observed may be due, at least in part, to the task or instructions that participants received.

We found that prefixed words were less transparent than suffixed words, a result that is coherent with previous studies conducted in French (Colé et al., 1989; Kandel et al., 2012). We also found that words that are both prefixed and suffixed are less transparent than words that are either suffixed or prefixed. The meaning of suffixed words would be accessed through the root, maintaining and reinforcing the relationship between the full complex word and its components, while prefixed words would be processed as simple words (Colé et al., 1989; Kandel et al., 2012). Words that are both prefixed and suffixed may be affected by non-additive transformation effects.

We did not find any significant effect of sociodemographic variables on semantic transparency ratings. Of all the variables that were examined (age, sex, education, learning history, and number of languages spoken as fixed factors), education had the strongest effect, albeit a nonsignificant one. It is important to acknowledge that most participants were highly educated, with only a small proportion lacking post-secondary education—a limitation for sample representativity that is well documented (Reinikainen et al., 2018). Studies focusing on the effect of education on semantic transparency ratings should include a more evenly distributed sample to avoid a restriction of range issue such as the one observed in our sample.

Speakers of French as a second language provided ratings which were similar overall to those of native French speakers. Looking more closely at the influence of individual French proficiency on ratings, we found that reading proficiency and oral comprehension proficiency had an influence on semantic transparency ratings. More specifically, higher ratings of reading proficiency were associated with smaller differences with the ratings of native French speakers. However, higher ratings of oral comprehension proficiency were associated with larger differences with the ratings of native French speakers. Although we do not have detailed language acquisition history and current experience and exposure data for speakers of French as a second language, our results support a general association between morphological processing and second-language proficiency (Brooks et al., 2011; Kimppa et al., 2019; Lam & Chen, 2018). Since words were presented in writing, it is not entirely surprising that semantic transparency ratings were influenced by reading proficiency, and that higher reading proficiency was associated with more similar ratings (see also Beyersmann et al., 2020). However, the result for oral comprehension proficiency is more surprising. It is possible that phonological information was more readily available for participants with higher oral comprehension proficiency. However, French is characterized by highly inconsistent grapheme-phoneme correspondence (Ziegler et al., 1996). Speakers of French as a second language may handle interference due to phonological/orthographic disparities differently than native speakers, especially if their native language has a shallow orthography (Abu-Rabia et al., 2013; Ramirez et al., 2010). Similarities and cognate status might also have influenced semantic transparency ratings in speakers of French as a second language (e.g., consider a native English speaker rating the morphologically related pair "barbe"—"barbier" while accessing "barb", "beard" and "barber", i.e., a cognate (barber), a non-cognate (beard), and a false friend (barb) that are not morphologically related in English) (Comesana et al., 2018; Commissaire, 2022; Kahraman & Kırkıcı, 2021; Ramirez et al., 2013). Our results are coherent with those of recent studies but would require formal testing and replication in a larger and more tightly controlled sample.

Research and clinical applications of SyllabO+

The updated SyllabO+ database of real oral Québec French has numerous applications for teaching and learning French, and in the field of speech-language pathology with individuals facing difficulties with oral or written language. Indeed, knowing syllables and morphological boundaries in words is important for both reading and spelling, and morphological awareness is related to second language proficiency in childhood and adulthood (Brooks et al., 2011; Kimppa et al., 2019; Lam & Chen, 2018).

Language learners and teachers can therefore use our databases (and our new online « dictionary ») to decompose words into their constituent morphemes and syllables. Decomposing words offer insights into a language's mechanics and meaning. For example, language learners can more easily learn that the morpheme –s signals the plural in writing, and that the ending of words that mean "to do something in a X way" (*rapide-ment* – rapidly, *facile-ment* – easily) is consistently written <ment>, which helps them distinguish it from homophonic endings (e.g., present participle: <ant>). Our database provides educators new materials to quickly develop morphological awareness exercises adapted to the Québec French variety starting from the syllable, the morpheme, the lexeme, or the word, depending on their focus. The frequency information that is included in the databases for all units of language (phonemes, syllables, morphemes, lexemes, and words) will help educators and speech-language pathologists in selecting among the most common syllables, words, prefixes, and suffixes of the contemporary Québec French or, alternatively, in choosing rare items to increase difficulty, allowing for the construction of exercises of various levels of difficulty and developmental levels.

This material can equally well be used in psycholinguistic or cognitive neuroscience of language research to assess various skills including (but not limited to) syllable word and nonword repetition (the database of syllable pairs and triads provides ample materials that can be used as nonwords), phonological manipulation, lexical decision, and speech perception at the lexical and sublexical levels including identification and discrimination tasks. Not only can researchers manipulate lexicality (words and nonwords), but they can also manipulate complexity (e.g., presence of a consonant cluster), frequency, and length (e.g., number of syllables).

Conclusion

In the present paper, we describe several major changes that were made to the SyllabO+ project: we included 41 additional speakers, and we conducted lexical and morphological analyses. These new data have been made available on our website (https://syllabo.speechneurolab.ca), which now also features a dictionary-style interface. This interface allows users to search for words and access information such as syllabic segmentation and spoken frequency for each syllable and morpheme, as well as additional linguistic metrics. While similar information has been made available for French in the past, SyllabO+ is unique in reporting indicators based on real spontaneous spoken language instead of written or scripted language (e.g., movies or movie subtitles). It is also unique in its inclusion of a large number of indicators from several domains of language (e.g., phonetics, phonology, lexicon, morphology) that were all extracted from the same talkers. This new version of SyllabO+ opens up new research perspectives and the possibility to revisit questions that were left unanswered. Both the corpus analysis statistics and data from the study on semantic transparency can be used by researchers to test other hypotheses related to first and second language proficiency, language acquisition, and speech/language impairment.

Acknowledgements We thank Alexis Piéplu for his help with the development, maintenance, and hosting of the SyllabO+ web application, and all the speakers.

Authors' contributions Noémie Auclair-Ouellet: Conceptualization, Methodology, Validation, Supervision, Project administration, Funding acquisition, Writing—Original Draft, Writing—Review & Editing. Alexandra Lavoie: Formal analysis, Writing—Review & Editing. Pascale Bédard: Methodology, Software, Writing—Review & Editing. Alexandra Barbeau-Morrison: Formal analysis, Investigation, Writing—Review & Editing. Patrick Drouin: Funding acquisition, Methodology, Writing—Review & Editing. Pascale Tremblay: Conceptualization, Methodology, Validation, Data Curation, Resources, Supervision, Project administration, Funding acquisition, Writing—Original Draft, Writing—Review & Editing.

Funding This project was supported by an *Insight Development Grant* from the Social Sciences and Humanities Research Council of Canada to NAO and PT (430-2018-00411) as well as an Insight Grant from the Social Sciences and Humanities Research Council of Canada to PT (435-2023-057).

Code availability Not applicable.

Declarations

Ethics approval The original study was approved by the *Comité d'éthique de la recherche sectoriel en neurosciences et santé mentale, Institut Universitaire en Santé Mentale de Québec* (#356-2014). The semantic transparency online study was approved by the Institutional Review Board of the Faculty of Medicine and Health Sciences at McGill University (#A04-E24-20B) and the *Comité d'éthique de la recherche sectoriel en neurosciences et santé mentale, Institut Universitaire en Santé Mentale de Québec* (#2021–2143).

Consent to participate All participants provided informed consent to participate.

Consent for publication All authors consent to the publication of this article.

Conflicts of interest/Competing interests None.

References

Abeillé, A., & Clément, L. (2003). *Les mots simples - Les mots composés Corpus Le Monde*. http://llf.linguist.jussieu.fr

Abu-Rabia, S., Shakkour, W., & Siegel, L. (2013). Cognitive retroactive transfer (CRT) of language skills among bilingual Arabic-English readers. *Bilingual Research Journal, 36*(1), 61–81.

Arnaud, P. J. L., & Renner, V. (2014). English and French [NN]N lexical units: A categorial, morpho- logical and semantic comparison. *Word Structure*, *7*(1), 1–28. hal-01097867

Ashkenazi, O., Gillis, S., & Ravid, D. (2020). Input-output relations in Hebrew verb acquisition at the morpho-lexical interface. *Journal of Child Language, 47*(3), 509–532. https://doi.org/10.1017/S0305000919000540

Baudot, J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain*. Les Presses de l'Université de Montréal.

Bauer, L. (2008). Derivational morphology. *Language and Linguistics Compass, 2*(1), 196–210.

Beauchemin, N., Martel, P., & Théorêt, M. (1992). Dictionnaire de fréquence des mots du français parlé au Québec: Fréquence, dispersion, usage, écart réduit. *Linguistics, 26*, 775. https://books.google.ca/books?id=mfnjAAAAMAAJ

Bédard, P., Audet, A. M., Drouin, P., Roy, J. P., Rivard, J., & Tremblay, P. (2017). SyllabO+: A new tool to study sublexical phenomena in spoken Quebec French. *Behavior Research Methods, 49*(5), 1852–1863. https://doi.org/10.3758/s13428-016-0829-7

Beyersmann, E., Mousikou, P., Javourey-Drevet, L., Schroeder, S., Ziegler, J. C., & Grainger, J. (2020). Morphological processing across modalities and languages. *Scientific Studies of Reading, 24*(6), 500–519.

Bowers, P. N., & Kirby, J. R. (2010). Effects of morphological instruction on vocabulary acquisition. *Reading and Writing, 23*(5), 515–537.

Brooks, P. J., Kempe, V., & Donachie, A. (2011). Second language learning benefits from similarity in word endings: Evidence from Russian. *Language learning, 61*(4), 1142–1172.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106–116.

Carlisle, J. F., McBride-Chang, C., Nagy, W., & Nunes, T. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly Journal of Experimental Psychology, 45*(4), 464–487. https://doi.org/10.1598/rrq.45.4.5

Casalis, S., Cole, P., & Sopo, D. (2004). Morphological awareness in developmental dyslexia. *Annals of Dyslexia, 54*(1), 114–138. https://doi.org/10.1007/s11881-004-0006-z

Cavalli, E., Duncan, L. G., Elbro, C., El Ahmadi, A., & Cole, P. (2017). Phonemic-Morphemic dissociation in university students with dyslexia: An index of reading compensation? *Annals of Dyslexia, 67*(1), 63–84. https://doi.org/10.1007/s11881-016-0138-y

Cole, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words - A differential frequency effect. *Journal of Memory and Language, 28*(1), 1–13. https://doi.org/10.1016/0749-596x(89)90025-9

Comesana, M., Bertin, P., Oliveira, H., Soares, A. P., Hernandez-Cabrera, J. A., & Casalis, S. (2018). The impact of cognateness of word bases and suffixes on morpho-orthographic processing: A masked priming study with intermediate and high-proficiency Portuguese-English bilinguals. *PLoS One, 13*(3), e0193480. https://doi.org/10.1371/journal.pone.0193480

Commissaire, E. (2022). Do both WRAP and TRAP inhibit the recognition of the French word DRAP? Impact of orthographic markedness on cross-language orthographic priming. *The Quarterly Journal of Experimental Psychology (Hove), 75*(6), 1094–1113. https://doi.org/10.1177/17470218211048770

Diependaele, K., Grainger, J., & Sandra, D. (2012). Derivational morphology and skilled reading: An empirical overview. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 311–332). Cambridge University Press. https://doi.org/10.1017/CBO9781139029377.021

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology, 9*(1), 99–117.

Durand, J., Laks, B., & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language* (pp. 93–106). Gunter Narr Verlag.

Elbro, C., & Arnbak, E. (1996). The role of morpheme recognition and morphological awareness in dyslexia. *Annals of dyslexia, 46*(1), 209–240. https://doi.org/10.1007/BF02648177

Fradin, B. (2011). IE Romance: French. In R. Lieber & P. Štekauer (Eds.), *Oxford handbook of compounding*. University Press.

Fradin, B., Dal, G., Grabar, N., Lignon, S., Namer, F., Tribout, D., & Zweigenbaum, P. (2008). Remarques sur l'usage des corpus en morphologie. *Langages, 3*, 34–59.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical education, 38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jared, D., Jouravlev, O., & Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(3), 422–450. https://doi.org/10.1037/xlm0000316

Kahraman, H., & Kırkıcı, B. (2021). Letter transpositions and morphemic boundaries in the second language processing of derived words: An exploratory study of individual differences. *Applied Psycholinguistics, 42*(2), 417–446.

Kandel, S., Spinelli, E., Tremblay, A., Guerassimovitch, H., & Álvarez, C. J. (2012). Processing prefixes and suffixes in handwriting production. *Acta Psychologica, 140*(3), 187–195.

Kimppa, L., Shtyrov, Y., Hut, S. C. A., Hedlund, L., Leminen, M., & Leminen, A. (2019). Acquisition of L2 morphology by adult language learners. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 116*, 74–90. https://doi.org/10.1016/j.cortex.2019.01.012

Lam, K., & Chen, X. (2018). The crossover effects of morphological awareness on vocabulary development among children in French immersion. *Reading and Writing, 31*(8), 1893–1921.

Lehmann, A., & Martin-Berthet, F. (2005). *Introduction à La Lexicologie : Sémantique Et Morphologie* (2nd edition ed.). Armand Colin. https://www.abebooks.com/9782200342999/Introduction-lexicologie-Sémantique-morphologie-Lehmann-2200342993/plp

Mailhot, H., Wilson, M. A., Macoir, J., Deacon, S. H., & Sanchez-Gutierrez, C. (2020). MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods, 52*(3), 1008–1025. https://doi.org/10.3758/s13428-019-01297-z

Marslen-Wilson, W., & Zhou, X. (1999). Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes, 14*(4), 321–352.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*(1), 3–33.

Martin, J., Frauenfelder, U. H., & Colé, P. (2013). Morphological awareness in dyslexic university students. *Applied Psycholinguistics, 35*(6), 1213–1233. https://doi.org/10.1017/s0142716413000167

Mathieu-Colas, M. (1996). Essai de typologie des noms composés français. *Cahiers de Lexicologie, 69*, 71–125.

Medeiros, J., & Dunabeitia, J. A. (2016). Not everybody sees the ness in the darkness: Individual differences in masked suffix priming. *Frontiers in Psychology, 7*, 1585. https://doi.org/10.3389/fpsyg.2016.01585

Namer, F. (2003a). Automatiser l'analyse morpho-sémantique non affixale: le système DériF. *Cahiers de Grammaire, 28*, 31–48.

Namer, F. (2003b). Le modèle Lstat : Ou comment se constituer une base de données morphologique à partir du Web. *Revue Québécoise de Linguistique, 32*(1), 85–109.

New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. Actes de la Conférence Traitement Automatique des Langues Naturelles, Louvain.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, *101*, 447–462. http://www.lexique.org

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*(2), 244–247. https://doi.org/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*(3), 674–685. https://doi.org/10.1111/j.1467-8624.2009.01290.x

Poirier, C. (2009). Le français d'Amérique: une variété maternelle distincte. *Québec Français*, *154*, 39–41. https://www.tlfq.org/publications/le-francais-damerique-une-variete-maternelle-distincte

Ramirez, G., Chen, X., Geva, E., & Kiefer, H. (2010). Morphological awareness in Spanish-speaking English language learners: Within and cross-language effects on word reading. *Reading and Writing, 23*(3), 337–358.

Ramirez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-Speaking English language learners: The facilitating effect of cognate knowledge. *Topics in Language Disorders, 33*, 73–92.

Reinikainen, J., Tolonen, H., Borodulin, K., Harkanen, T., Jousilahti, P., Karvanen, J., Koskinen, S., Kuulasmaa, K., Mannisto, S., Rissanen, H., & Vartiainen, E. (2018). Participation rates by educational levels have diverged during 25 years in Finnish health examination surveys. *European Journal of Public Health, 28*(2), 237–243. https://doi.org/10.1093/eurpub/ckx151

Rey-Debove, J. (2004). *Le Robert brio: Analyse comparative des mots*. Éditions Le Robert.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. http://www.ncbi.nlm.nih.gov/pubmed/8943209

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*(1), 27–52.

Séguin, H. (1993). Fréquences d'utilisation des mots en français écrit contemporain. *Revue québécoise de linguistique, 22*(2), 179–181.

Singson, M., Mahony, D., & Mann, V. (2000). The relation between reading ability and morphological skills: The evidence from derivational suffixes. *Reading and Writing, 12*, 219–252.

van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38*(4), 584–589. https://doi.org/10.3758/bf03193889

Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, & Computers, 28*(4), 504–515.

Open practices statement The data and materials for all experiments are available. The corpus (XML transcription) and all databases are freely available on the SyllabO website at https://syllabo.speechneurolab.ca as well as on Borealis, the Canadian Dataverse Repository, at https://doi.org/10.5683/SP3/T3ZUIN. The original voice recordings cannot be shared because at the time the recordings were made, participants did not consent to public data sharing of their voice.

The work detailed in this article was not preregistered.