DOI: 10.1111/modl.70021

The Modern Language Journal

ORIGINAL ARTICLE

A nuclear families word list for French

Thomas Cobb[1] | **Christina Lindqvist**[2] | **Mårten Ramnäs**[2]

[1]Didactique des langues, Université du Québec à Montréal, Montreal, Canada

[2]Department of languages and literatures, University of Gothenburg, Gothenburg, Sweden

Correspondence
Thomas Cobb, Didactique des langues, Université du Québec à Montréal, 1205 rue St Denis, Montreal, Canada.
Email: cobb.tom.3@gmail.com

Abstract

This between-languages replication study relates the development and testing of a nuclear, families-based, pedagogical word list for French as was previously done for English. A word family includes base and inflected words (or lemmas) plus derivations. A nuclear family is reduced to the most frequent of these, with less frequent members set aside according to a frequency criterion. Such a list is needed in French because existing French word lists are impractically large in size yet insufficient in coverage and being lemma based obscure the relationship between base, inflected, and derived words. The base list for nuclearization in French was created through a process of finding, fleshing out, and familizing a lemma list. The resulting base list is 3000 word families (25,141 word types) with 96%–98% coverage across a range of text types, in itself unique in French pedagogy. Nuclearizing the base list involved deriving one or more sublists from the base list through two user choices, a corpus representing a particular topic of interest or level of study and a frequency criterion within that corpus for inclusion of base list word types. The resulting nuclear list is 2,871 families in 10,458 word types with ~90% coverage in texts potentially used in reading instruction.

KEYWORDS

derivational morphology, French as a foreign language (*français langue étrangère*), lexical coverage, pedagogical word lists, usage-based language learning, word counting units

It would be hard to overestimate the role of frequency lists in the expansion and enrichment of vocabulary research and teaching in English as a second or foreign language (ESL/EFL) since about 1980. Lists are at the heart of text profiling, materials grading, course sequencing, text simplification, and

vocabulary testing, the activities that define vocabulary instruction as we know it today. The lists in question are not those of the audiolingual era, with words selected to highlight grammar structures and otherwise keep out of the way, nor of the communicative era, serving up the keywords of particular life schemas (ordering in a restaurant and reading a bus timetable) or reading themes (how windmills work, navigation in bats) that were unlikely to be seen again. The modern corpus-based, frequency-structured word list is based on different principles. It focuses attention on high- and medium-frequency words, that will definitely be seen again, and that will become the basis for learning less frequent and specialist terms. It is more likely to inhabit a computer program than the back pages of a textbook or workbook. The learners who use it may never actually see the list (though that is an option) but rather meet texts and other inputs informed by it.

The pinnacle of list work in English to date is Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) word list of 25,000 word families, based on both the BNC (Oxford Computing, 1995) and the COCA (Davies, 2008). This list is the beneficiary of more than 100 years of pedagogical list building in English. It is the most complete pedagogical word list ever made, with the highest overall coverage ever reached across a range of text types and both "liked" and judged "useful" by English teachers (Dang & Webb, 2020). The BNC/COCA is employed by numerous researchers, teachers, and learners for a wide range of purposes, especially through its application in list-driven text profiling software like *AntwordProfiler* (Anthony, 2024; https://www.laurenceanthony.net/software/antwordprofiler/) and *Vocabprofile* (https://lextutor.ca/vp/comp/). Similarly, *MultilingProfiler* (Marsden et al., 2023; https://www.multilingprofiler.net/) has recently been developed for French, German, and Spanish. Profiling is the classification of every word of a text according to its membership in a set of frequency lists. It allows teachers to find, create, or modify texts in terms of overall lexical frequency, for example, to reduce a text to 98% high frequency words for lower intermediate readers (or roughly A2 in the Common European Framework of Reference for Languages [CEFR]; Council of Europe, 2020).

Despite the success of modern list work in English, however, there remain usability problems with even the best lists. The BNC/COCA achieves its high coverage mainly through its sheer size. The list comprises 25,000 families, which, with an average of 5–7 word forms per family (Nation, 2006), amounts to between 125,000 and 175,000 different words, or word types. Even the lower estimate is the size of a typical novel (*Lady Chatterley* is 121,136 words). The advantage of a list this large is that it contains almost any word a learner will ever come across; the profiler employing it will rarely return "not found." The disadvantage is that a teacher cannot use the list directly with learners. A teacher cannot give a segment of a raw list to learners to work through, say, building a glossary of unknown words, as described in Cobb (1999): Just the first 3000 BNC/COCA word families involve 19,062 word types with no clues which to prioritize. It is thus more common to give learners just head words (base words) to work with in such projects, but doing this prevents them from meeting and processing the inflected and derived forms that are the majority of the word's employments in natural language (the infinitive form of the French verb "venir [to come]" is only 5.7% of the verb's occurrences in the literary subcorpus of Chambers & Le Baron's, 2006, corpus of French research articles).

For reasons like these, Cobb and Laufer (2021) looked for a way to reduce the BNC/COCA to smaller subsets of high-frequency items for particular levels and purposes. On the success of this approach, the present researchers now propose to follow a parallel process for the development of a French list, though not expecting the parallels to be more than starting points. Our goal is to see if it is possible to give French teachers, course designers, and testers a way of providing a principled lexical basis for French instruction wherever French is taught at any level from lower intermediate to advanced (or CEFR A2, post-beginner, to C1, advanced). We focus on French as a foreign language (*français langue étrangère* [FLE]) rather than French as a second language (*français langue seconde* [FL2]) in order to isolate the effects of instruction per se rather than mixed with those of living or studying in French.

LITERATURE REVIEW

This section has three parts: (a) an account of why and how a nuclear list was built in English, (b) why a new word list is needed in French, and (c) why the list should be a nuclear families list, including anticipation of both technical and pedagogical challenges with particular reference to the incorporation of derived words within families.

Making word lists

The science of list building (e.g., Dang, 2020; Nation, 2016) and list evaluation (e.g., Dang & Webb, 2020) has advanced rapidly in recent years. Word lists can be intuition based or single-text based (like glossaries at the end of stories), but the lists of interest here are corpus based, drawn on a large collection of texts. Like dictionaries, corpus lists were previously made by hand (*General Service List*, 1953; *Le Français fondamental* [Fundamental French], 1956) with a support technology of cards, boxes, and random informants, but more recently have relied on text computing and particularly computers' ability to count, track, assemble, and list character strings in large, sampled text collections called corpora. A corpus is collection of texts, usually divided into subcorpora, and can be domain general (*Brown*; Kucera & Francis, 1967, in English) or domain specific (*Scientext*; Tutin & Grossman, 2014, in French). One million words was once seen as large in corpora (*Brown*), while nowadays 10 billion is common (*TenTen*; Jakubíček, Kilgariff et al., 2013).

As early as the 1960s, computers could extract, list, and count all the words in a corpus individually, but because this led to enormous lists of single words, some sort of grouping was normally employed. The simplest grouping is by word types, where, for example, one finds "walk_3" rather than "walk, walk, walk," though a word-types list for a corpus of any size is still enormous. Two further grouping refinements can reduce list size further, and in ways that potentially model how words are stored in the mental lexicon. Specifically, words can be grouped by lemmas ("walk," "walked," "walks," "walking," etc., within the same part of speech [POS], all verbs), or families (lemmas plus related words in different POSs, like "walk" [verb] and "walker" [noun]). In terms of size reduction, family grouping provides the greatest reduction in the number of units. However, despite the progress of recent years, an abiding limitation to the value of word lists in language learning is their size in whatever unit.

List nuclearizing in English

In view of the inevitable trade-off between comprehensiveness and utility in word lists, Cobb and Laufer (2021) developed a rationale and technology for systematically focusing subsets of large lists on particular levels or topics through a reduction process called nuclearization. Nuclearization identifies the extent to which the lemmas and derivations of particular head words occur in a chosen corpus, thus providing a principled way of selecting the most frequent of these, or the nucleus, for emphasis during the learning process. An example is the French family "charger [to charge]," which has 87 members in 632 occurrences in a corpus of learner materials but only 11 of the 87 account for 607 of the occurrences; these 11 more frequent members form a plausible nucleus. Cobb and Laufer nuclearized the first 3000 families of the BNC/COCA (a list of 19,062 word types providing ~90% coverage in intermediate and higher reading materials) as follows:

1. For a hypothetical learner group of pre-academic intermediate ESL learners, a comparison or cross-corpus of over 2 million words of typical reading materials was assembled.
2. Software was developed to count the number of occurrences of each BNC/COCA/3000 word type in the cross-corpus and calculate these as a percentage of occurrences in its family.

3. The within-family percentages served as cutoff points for inclusion in the generation of sublists from the base list (e.g., family members were included that were more than *x*% of their families).
4. The best cutoff was determined by generating lists for a range of cutoffs and comparing them for (a) size and (b) coverage in a new but similar corpus (on the assumption that the best list is the smallest list with the most coverage).

The finding was that a list of words that represented 7% or more of their family composition in the cross-corpus produced a nuclear list of just 7293 word types (originally 19,062) with a coverage of 84.5% (originally 90%), a "good trade" in cost–benefit terms.

In addition to showing that a nuclear list greatly reduced learning burden without seriously compromising coverage, the work by Cobb and Laufer (2021) revealed something unexpected: Nuclear families proved almost always to contain at least one high-frequency derivation. For example, "arrange" has 38 members, but when nuclearized at 7% has just 5 members, one of which is the derived from "arrangement." The complete family buries "arrangement" amid less frequent derivations like "prearrange," "arrangers," and nine others where it is unlikely to be noticed. The researchers reasoned that the increased salience of derivations could be expected to promote the association and integration of derived words into their families in learner lexicons. In French, we hope to find similar reduction opportunities, coverage advantages, and derivation highlighting.

Does French need another vocabulary list?

From *Le français fondamental* (Gougenheim et al., 1956) to a new examinations list from the United Kingdom Department for Education (Finlayson & Marsden, 2024) there has been no shortage of word lists for the *didactique du français* [French teaching]. Is a nuclear list just one more list thrown into the fray? We propose that a nuclear list solves problems that no list has yet solved.

Despite their existence, French word lists have not always been extensively used and have not typically been central to French teaching. French teaching has traditionally focused on the early development of form-focused accuracy in productive skills at the expense of both spoken communication and receptive skills like listening and reading. Indeed, one of the goals of the postwar CEFR reforms—a "profound reorientation" (Trim, 2012, p. 22)—in language teaching was to correct an imbalance, in France and other participating countries, in traditional models of instruction that stressed "excessive formalism" (p. 25), that is, in spoken and written production, in favor of a more balanced development of interactional competencies "across a range of skills and use" (p. 29) including receptive skills like listening and reading.

The point of word lists historically has been to serve listening and reading development, since these skill areas can present learners with unanticipated vocabulary; learners meet a glossary when preparing for a listening passage, or at the end of a reading passage, not at the end of a grammar exercise or the beginning of an oral presentation. We position our list work within the ongoing CEFR reorientation, while taking up the challenge that Trim and other members of the original project were not list supporters. They had decided "not, at that time [to attempt] constructing a scale of ascending levels defined in terms of frequency of vocabulary" (Trim, 2012, p. 25) for the following reasons: There were no lists based on spoken corpora at the time, while CEFR emphasized spoken communication; CEFR aimed at a pan-European system with several languages, many of which had no corpora to make lists from; and, worst, adherence to common word lists would amount to a top-down model that would be at odds with the goal of "empowering teachers to plan courses as close to the point of learning as possible" (Trim, 2012, p. 35).

To this day there remains no official CEFR word list. There is no shortage of interest in one, at least for English, as shown, for example, in the *Englishprofile* website (https://englishprofile.org/?menu=evp-online), where one can profile a text according to a set of purported CEFR lists, produced by Cambridge University Press, but not access the lists themselves. There has been no lack

of funding or expertise available for such a list, for example, for the ambitious, methodologically impeccable, European-Union-funded 'Kelly' project (Keywords for Language Learning for Young and Adults Alike) of roughly 2009–2011, "the main objective of which was to create vocabulary lists for nine languages (Swedish, English, Norwegian, Greek, Italian, Polish, Arabic, Chinese, and Russian) and adapt them to CEFR levels" from A1 to C2 (Kokkinakis & Volodina, 2013, p. 211). Note, however, that French and German were not among the languages, and only the Swedish list was ever officially made available (though others can be accessed at a Github site run by Sharoff at https://ssharoff.github.io/kelly/). The Kelly project was followed by another: the CEFRLex project (https://www.uclouvain.be/fr/instituts-recherche/ilc/cental/cefrlex), which, among others, contains a French list (called FLELex, described in François et al., 2014). The 47,432 lemmas of FLELex are graded A1 through C2 based on a small corpus of texts that had been graded by CEFR levels, raising the question of how the levels of the texts were established without a list. Perhaps for this reason the CEFRLex lists are not recognized as official within CEFR. The lack of an official, available, and easily usable, word list contributes, arguably, to a general lack of specification within the CEFR approach, that testers like Alderson (2007) have noted creates problems for measuring attainment within CEFR programs, especially in "reading and listening" (p. 66). We believe the nuclear approach to list construction made possible by today's more powerful computers, accessible corpora, and tech-savvy practitioners can overcome these limitations and produce simple, usable lists for any language that has a corpus. We return to CEFR integration in the discussion.

The main problem a word list would address is that many or most FLE learners do not acquire sufficient vocabulary from their school studies to do anything interesting in French, including pursue further studies in French or in French medium. This was long suspected, though not shown empirically, possibly due to the lack of appropriate word lists on which vocabulary measures could be based (as Nation & Beglar's, 2007, Vocabulary Size Test is based on the BNC/COCA). Awaiting better lists, however, researchers have used the lists they had to put together a critical mass of studies of both learners and the input they receive. Using a test based on head words from Baudot's (1993) lemma list, Milton (2006) assessed the lexical acquisition of British FLE school learners who had no exposure to French outside school and had not been given specific vocabulary training. He found that after 5 years of French instruction, such learners' average receptive lexicon was "between 800 and 1,000 words of French" (Milton, 2006, p. 193). Similar findings are provided by David (2008) in a British context, De La Maya Retamar and Mora Ramos (2019) in a Spanish context, and Lindqvist (2018) in a Swedish context. Milton further determined that the few French words the British learners did know had been acquired in the first year, with little learning thereafter. Is there any clue to this pattern in the learning materials these learners are offered? (Any such pattern would of course be correlational, not necessarily causal.)

The lexis of French textbooks

We have found three corpus-oriented studies of textbooks used in FLE. In the first, subtitled "Do learners have a chance?" Tschichold (2012) looked at a 4-year series of French textbooks, *Encore Tricolore* (Mascie-Taylor & Honnor, 2001), in use across UK schools in the period of Milton's work. Her analysis compared the vocabulary presented in *Tricolore* to the 1500 lemmas of *Français fondamental, 1er degré* [Basic French, Level 1] (FF1; further discussed below, a lemma-based pedagogical list developed for French in the 1950s which Tschichold equated to the B1 level of CEFR). *Tricolore* presents learners with 2656 lemmas altogether, so Tschichold's interest was more in how these words were chosen and sequenced, and whether the most useful words were actually included, than in brute quantity. Her finding was that learners were exposed to just a small, easified portion of the FF1, with 50% fewer verbs and verb-derived words than were actually present in FF1, and a higher proportion of English cognates. Particularly interesting is the general lack of derived words in *Tricolore*, despite

FF1's sizeable collection. And, for the words presented, there was a low degree of recycling from unit to unit, especially for Books 2 and 3, shedding light on Milton's finding that most learning happens in the first year. Tschichold concluded that *Tricolore* users were unlikely to "reach level B1 of CEFR" (p. 7).

Milton and Hopwood (2022) looked at a second UK series, *Studio* (Bell & McLachlan, 2012), which was explicitly designed to update *Tricolore Encore* and prepare learners with examination-ready vocabulary levels. In a corpus methods analysis, they found that instead of increasing the vocabulary learning opportunities of *Tricolore*, *Studio* had reduced them by 40% (from 2656 to 1482 words available for learning). Learning opportunities were particularly few in the second volume onward, so again most of any learning will happen in the first year. It is as if, after presenting some basic vocabulary in volume one, textbook writers are not certain where to go next. Indeed, without a frequency list as guidance, this is likely to be the case. Perhaps relevant here is McCrostie's (2007) study of ESL professionals' frequency intuitions, which he found to be weak, particularly in the key mid-frequency zone (Schmitt & Schmitt, 2014), that is, at other than the high and low ends of the frequency spectrum. Milton and Hopwood (2022) attribute these inadequacies in *Studio* to the lack of "a clear standard for textbook writers to work to" (p. 667), that is, a word list.

Such tendencies are not confined to textbook production in the British Isles. Arndt (2025) performed a corpus analysis of a series widely used in the United States, the four-volume *T'es Branché* (Theisen, 2019) for Grades 6–12. As in the previous studies, the researcher found that the first 1000 words were fairly well represented, with sufficient recurrence and recontextualization to support learning, in principle, but that few of the words from the second and third 1000 zone recurred more than once or twice. Again, the point is made that post-1000 words will not make it into a syllabus except with the use of a vocabulary framework, that is, a word list that goes beyond 1000 lemmas.

What can a learner do with 1000 lemmas?

Reading a text with comprehension is normally shown to depend mainly on knowing the meanings of most of its words (Stæhr, 2008), which Laufer and Ravenhorst-Kalovski (2010) specified as knowing 95% of them. Text analysis with Vocabprofile (https://lextutor.ca/vp/comp/) shows that 3000 French lemmas are needed to reach 95% coverage in an article from Montreal's *Le Devoir*; 4000 lemmas in Camus' (1942) *L'Étranger*; and 6000 in an article from *Le Monde Diplomatique*. Or, if a bare minimum of comprehension is the goal, which Laufer (2020) calculated is possible with knowledge of 90% of the words, *Le Devoir* and *L'Étranger* both demand 2000 lemmas to reach that figure, *Le Monde Diplomatique*, 3000, even if proper nouns are thrown in as "known." It seems fair to say that few FLE graduates can make much sense of any of these texts in return for their 4–5 years of study.

Speaking and listening presents only a slightly better picture. In 45,000 word tokens from Beeching's (1997) corpus of conversational French, knowledge of 1000 lemmas provides 89% coverage, such that, with the extra contextual support of the spoken medium, some learners might reach basic competence with about 1000 lemmas, provided they knew the pronunciations.

Challenges of developing a nuclear list in French

Technical challenges

The main challenge of nuclearization in French is that there is no comprehensive family word list to start from that is comparable to the BNC/COCA. Most frequency-based French word lists come out of corpus linguistics rather than pedagogy, and are lemma based; English word lists are more commonly produced by or in consultation with language educators and are family based. The BNC and COCA

individually were produced with direct input from pedagogy (Gardner & Davies, 2013), and of course Nation's BNC/COCA blend is entirely pedagogical in its goals and character.

A key characteristic of a pedagogical list is that derived words are grouped with their base words, where they are likely to be noticed and linked in teachers' and learners' minds, whereas a lemma list counts derived forms as separate lemmas that can be spread out over several regions of a list. In one lemma list we shall look at, that of Lonsdale and Le Bras (2009), both the adjective form "possible [possible]" and noun form "possibilité [possibility]" are found in the first 1000 items, albeit as separate items, while "impossible [impossible]" is in the second 1000 and "impossibilité [impossibility]" in the fifth. If a lexical syllabus were to be built on such a word list, "impossibilité" would not be met until late in the sequence, if ever, despite the fact that it is rather similar in form to "possibilité."

Whatever uses the lemma counting unit may have in corpus linguistics, it has also come to be used in word lists designated "pedagogical" in French. The Kelly and CEFRLex lists discussed above were not only lemmas but POS parsed lemmas. The Lonsdale and Le Bras lemma list referred to earlier is subtitled "Core vocabulary for learners." A new list called *Eduscol* developed by the French *Ministère de l'éducation nationale et de la jeunesse* [French Department of national education and youth] (2020) is lemma oriented ("voler [to fly]" and "vol [flight]" are separate entries). The current *Petit Robert* online *Dictionnaire d'apprentissage du français pour les étudiants de français langue étrangère* [Learner dictionary for students of French as a foreign language] (https://lerobert.com/dictionnaires/fle/) is entirely lemma based. Within its entries, the extensive lexical resources of *Le Grand Robert* can be mined for synonyms, antonyms, and conjugations, but not for closely related derived forms, even where these are nearly identical to their base words, are high-frequency words in themselves, or are parts of high-frequency expressions.

Do classroom French teachers feel the need for a families list? Our impression is that many do but their views are rarely published. One that was published is Antes (2023), who notes the general lack of progress on vocabulary in US French instruction, arguing that only with an agreed-on word list can vocabulary learning targets be established and sequenced or fair assessment measures developed. She specifically proposes the development of a French version of West's (1953) *General Service List* (*GSL*), the original 2000-families list that was the point of departure for Nation's work in English. While we do not endorse that particular solution, the present work is performed in sympathy with this goal.

Are there really no family word lists for French? There have been dictionaries based on word families, such as Davau et al.'s (1972) *Dictionnaire du français vivant* [Dictionary of Modern French] whose definitions included subentries of both inflected and derived words, but it is no longer in print. The French version of the online WordReference dictionary supplies derived forms in many though not all of its French definitions (e.g., https://www.wordreference.com/fren/manger includes both verb and noun senses for "manger": "to eat" and "something to eat, food") and in any case does not supply its word lists. Gala and Rey (2008) developed a website called "Polymots" for French teachers that returns complete families for single word inputs but does not give access to its lists.

The only current French word list project we have found that claims to be family oriented is one produced for the UK Department for Education by Finlayson and Marsden (2024) for use in the development of the General Certificate of Secondary Education (GCSE) French syllabus and examinations. However, on inspection, the list is only nominally a families list. Its families are basically lemmas integrating a subset of derived words that meet certain conditions, as specified in an official document (https://www.gov.uk/government/publications/gcse-french-german-and-spanish-subject-content) which states: "If derived forms are used in listening or are required for production [in examinations], they will be listed separately in the Vocabulary List" (p. 29)—that is, not with the rest of their families—and further, that the list includes derived words only whose affixations "resemble English equivalents" and are "used in an English way" (p. 29). For example, words bearing "–ment" (roughly "–ly") appear in the list "only where the English equivalent is –ly" (thus excluding high-frequency items like "stationnement [parking]," "enseignement [teaching]," and "renouvellement [renewal]"). Such stipulations are clearly intended to assure lexical control of examinations, but

in so doing they render the list unusable for other pedagogical purposes. Its coverage in novel texts would be artificially reduced, though in any case coverage cannot be calculated since doing so would involve numerous human judgments. Also, the list cannot be used by the many French teachers and learners whose first language (L1) is not English.

To summarize, there is no corpus-based, family organized French word list that is equivalent to the first 3000 families of BNC/COCA that can serve as a comprehensive pedagogical list either in itself or as a basis for nuclearization.

Learning challenges

The development of a families-based word list mainly involves the incorporation of hitherto mid-frequency derived words into their related high-frequency lemmas. But does incorporating these derived forms within lemmas just shift the learning burden from more lemmas to more learning within lemmas (now families)? Applied linguistics research in English suggests that learners from some L1 backgrounds have difficulty recognizing or interpreting derived forms (e.g., McLean, 2017, though research with the same learners by Iwaizumi & Webb, 2022, finds the problem disappears with proficiency). In any case, we have seen that both textbook writers and GSCE list creators have misgivings about how French learners will cope with derived forms, so the question should be addressed if these are to be included in a pedagogical word list.

Here is how we foresee the learning of derived forms taking place by users of our list: Some of it can be reasonably predicted to happen by itself or with a minor nudge from a teacher. First, just putting derived words in physical proximity of the rest of their families in an in-use word list can be expected to create associations between base, inflected, and derived words, compared to keeping them apart in time and space as lemmas. Second, for the vast majority of FLE learners (with English or Romance language L1s), the most frequent French affixes are identical or similar to those of their own languages in both form and usage; English learners already know that "pre–" and "re–," or "–ment," and "–eur"—and probably a few more—are affixes, while Romance language learners probably know all of them. Third, research by Duncan et al. (2009) shows that the French derivational system is more easily learned than the English by L1 children in their respective languages because it is more regular and consistently applied. For example, French has almost no bound morphemes attached to derivational affixes, as abound in English (e.g., re+ceive, where *ceive is not standalone). These are possible reasons that both studies and training manuals for *la conscience morphologique* [morphological awareness (MA)], are relatively difficult to find for French as a foreign language, compared to English.

Despite such potentially facilitating factors, there are researchers in French instruction who describe the acquisition of derived words as a problem for learners. Fejzo (2020) described the derived word component of French as, while rather important (388 suffixes and 136 prefixes or 524 affixations, which form 75%–80% of French lexis, qua types though not tokens) also rather difficult, because derived words, individually, tend to be rare, and developing a training program in their adequate use over the course of school learning remains "in need of further study" (p. 8). We believe the number of current and frequent in-use affixes of French is almost certainly much lower than 524; that the supposed rarity of derived words is an artifact of adopting a lemma approach that counts derived words separately; and that the lack of success of morphology instruction reflects a bias toward production, particularly *l'orthographe grammaticale* [grammatical spelling], rather than comprehension (note that the classic MA studies in English, e.g., Kieffer & Lesaux, 2007, focus on receptive knowledge for reading comprehension).

We return to the question of learning derived words in the fourth research question, and in the discussion section we hope to provide a comprehension-based approach to morphology instruction that follows from our findings. Whether derived words present a problem to all FLE learners or not, clearly there is an argument for making the task more doable by identifying the words and affixes that

are the most important to teach and learn. With this goal in mind, we address the following research questions.

RESEARCH QUESTIONS

RQ1. Which existing French pedagogical list provides the best coverage across a range of text types?
RQ2. How much can the size of this list be reduced through familization (incorporating derived forms into families) and nuclearization (eliminating less frequent family members) without loss of coverage?
RQ3. What degree of reduction provides the smallest list for the greatest coverage?
RQ4. Does incorporating derived forms into families impose an unreasonable learning burden?

Each RQ will include its own methodology and results sections.

RQ1: METHOD

Base list selection

Our first goal was to find the best and most complete French list available to act as a basis for nuclearization. We first assemble a set of five candidate lists that resembled the BNC/COCA inasmuch as: they were either designated "pedagogical" or had been used at some point in education research (indeed in studies already cited); they were computation ready without requiring extensive preprocessing (i.e., to remove metalanguage or usage stipulations); and they did not involve POS parsing (which creates enormous numbers of head words unnecessarily). The retained lists were as follows:

1. Gougenheim et al.'s (1956) *Français fondamental, 1er et 2e degrés* (*FF*) comprising 3084 lemmas based on a spoken language corpus with a selection of texts.
2. Baudot's (1993) *Mots en français écrit contemporain* (its first 5000 lemmas) based on one million words of mainly literary and journalistic texts.
3. New et al.'s (2004) *Lexique 3* (first 5000) based on the 268-million-word *Frantext* mainly literary corpus, supplemented by the 50-million words SUBTLEX corpus (Brysbaert & New, 2009) of movie and television subtitles.
4. Lonsdale and Le Bras' (2009) *Core vocabulary for learners*, 5000 lemma headwords based on a corpus of 23 million words incorporating spoken and written French from numerous sources.
5. The French Department of National Education's (2020) *Eduscol*, a list proposed for use in primary schools based on a corpus of written, largely literary, French of the 19th and 20th centuries.
6. The British Department for Education's (2022) *Eduqas list* (discussed in Finlayson & Marsden, 2024), for use in the GCSE French syllabus, described as "classroom ready" in 2024 and "testworthy" in 2026; primarily derived from the 2000 most frequent headwords in Londsdale and Le Bras as well as additional frequency lists from a collection of young people's literature and previously used French examination scripts and familized as already described.

What would make one of these lists the "best" for our purposes? It would be the list that, despite being of a usable size, had the highest "coverage" in relevant learner texts or other learning materials, that is, whose words were most similar to the words in the materials.

What is a usable size? The size of a word list should reflect the number of words learners are expected to learn for receptive use. In one of the few studies of learning rate, Milton and Meara (1995) found that anglophone school-age French learners in a study year abroad learned receptively an average 550 headwords. Milton (2006) later found a similar figure for in-country learners preparing

French A Levels (i.e., advanced study). At a rate of 550 per year, for 5 years of instruction, learners would know 2750 headwords. We therefore reason that 3000 headwords, fleshed out as word families, is a useful first guess at the size of word list that could be used for 5 years of instruction, provided it offered acceptable coverage.

What is acceptable coverage? Laufer and Ravenhorst-Kalovski (2010) determined that 95% coverage would normally facilitate basic comprehension of age-appropriate materials if using resources (dictionaries and groupwork) while 98% was necessary for reading independently. These percentages were found to correspond to knowing 4000 families for 95% and 8000 for 98%. In view of the unlikelihood of all learners reaching these levels, Laufer (2020) kept reworking the problem eventually finding a basis for a 90% absolute minimum, corresponding to knowledge of as few as 3000 word families. In this study we will employ 90% as a minimal coverage figure.

Coverage contest

Before coverage and usability of these lists could be calculated and compared, list size had to be made as equal as possible and in several cases headwords fleshed out as lemmas, that is, with the addition of inflected forms. Approximate list size equality was achieved by reducing all lists to a maximum of the 5000 most frequent lemmas. Why 5000 lemmas if our goal was 3000 families? This is because familizing these lemmas, to follow in RQ2, will inevitably mean a reduction in the number of lemmas. For example, three related first 1000 level lemmas, "étude [study (n)]," "étudiant [student]," and "étudier [to study]," will be reduced to one family. In other words, 5000 lemmas were a guess at the number needed to assure 3000 families. Admittedly, using 5000 lemmas as a criterion created an inequality, as some of the lists contained fewer word types than that, but it should be remembered that small size was our other desideratum and small lists can achieve high coverage (e.g., the English nuclear lists, or Brezina & Gablasova's, 2013, *New General Service List*).

Expanding headword lists to lemma lists was achieved by raiding *Lexique 3*'s first 50,000 lemmas for its headword–inflection pairs and then integrating these into headwords lists, which made the lists significantly larger. For example, fleshing out Lonsdale and Le Bras's 5000 headwords in this manner yielded a collection of 36,945 word types, which may seem large until we remember that these figures will subsequently be reduced through nuclearization. Similar operations were performed for the *FF*, Baudot, *Eduscol*, and *Eduqas* lists, with corpus sizes in both lemmas and word types shown in Table 1.

Comparison methodology

Which existing French list of those we have looked at provides the best coverage across a range of text types and how many word types does it contain? To answer this, we developed a set of four mini corpora representing the types of texts different types of learners could arguably meet in an FLE course (chosen on the basis of long experience in FLE work). These are the four corpora:

1. 130,000 words of children's stories, a collection of the five main works of Goscinny's (1960) *Le Petit Nicolas*.
2. 100,000 words of literary works (Camus's novels other than *L'Étranger*, joined by two more recent novels: Modiano, 2001; Carrère, 2000).
3. 100,000 words of economics from Chambers and Le Barron's (2006) *Corpus of Research Articles in French*.
4. 100,000 words of unscripted conversational speech, drawn equally from the *Corpus de Français Parlé Parisien, CFPP* [Corpus of spoken Parisian French] (Branca-Rosoff et al., 2012) and the *Corpus de Français Parlé au Québec*, CFPQ [Corpus of spoken Quebec French] (Dostie, 2015).

TABLE 1 Sizes and coverages of French lemma lists.

List and year	Number lemmas	Number word types	Coverage (%) of lists by corpus				
			Children's literature	Adult literature	Academic/non-narrative	Unscripted speech	Mean coverage (*SD*)
FRANCAIS FOND, 1956	3084	23,532	63.58	65.12	58.44	61.46	62.15 (2.89)
EDUSCOL PRIMARY, 2020	1498	15,240	87.58	90.65	69.93	85.23	81.93 (10.74)
GCSE/EDUQAS (UK), 2024*	1621	9846	86.58	87.7	71.5	87.2	82.13 (9.21)
BAUDOT, 1993	First 5000	35,388	91.21	95.29	88.77	90.67	91.57 (3.35)
LEXIQUE 3, 2004	First 5000	32,744	92.70	96.12	86.53	93.65	92.10 (4.97)
LONSDALE & LEBRAS, 2011	5000	36,945	89.45	95.09	89.73	93.53	92.78 (2.75)
Mean (*SD*)		35,025.00 (2123.80)	85.18 (10.82)	95.47 (0.58)	88.33 (1.62)	92.60 (1.67)	

Abbreviation: GSCE, General Certificate of Secondary Education.

*As the GCSE list is under development, we have used a publicly available 2025 version of it that we found at https://doi.org/10.2307/3588328 (see Online Supporting Information) stripped of metalanguage and reduced to unique word types.

For each list, coverage was calculated as the number of times each word in the list appeared in each mini corpus, summed and expressed as a percentage of the total number of words in the corpus.

RQ1: RESULTS

Table 1 gives size details about each of the six contending lists and shows their coverages as percentages for each of the four text types. It shows the name and date of each list, its word count in types, and lemmas.

The data in Table 1 confirm several points about corpora, representativeness, and coverage that come straight from Corpus Ling 101: A quasicorpus list (*FF*, first row) does not achieve high coverage in any category, even its stated target of stories for children. Its smaller size may play a role, though the smaller but corpus-based *Eduscol* and *Eduqas* (second and third rows) are strong in coverage in the narrative texts largely used in schools. *Eduscol*'s relative weakness in unscripted (adult) speech and nonnarrative or academic text befits its designation as a primary school list; the similarly small *Eduqas* secondary school list has coverage of just under 90% in literary and a bit less in spoken texts, impressive for its small size, but just 71% in academic texts, reflecting its designation as a secondary, not university, word list.

The two penultimate lists (*Baudot* and *Lexique*) also show weakness in nonnarrative texts (88.77% and 86.53%) probably thanks to their literary origins, though strengths in speech, especially *Lexique* at 93.65%, probably reflecting its Subtlex component (transcribed speech). *Lonsdale and Le Bras*, the final list, has the highest coverage across text types, averaging 92.78%, with the lowest variability, $SD = 2.75$, probably reflecting its broad range of subcorpora and its size. Given these findings, the lemmatized *Lonsdale and Le Bras* will serve as the base for our nuclearization experiment.

We thus turn to RQ2, which asks: Can the size of *Lonsdale and Le Bras* be reduced through familization and nuclearization without significant loss of coverage?

RQ2: METHOD

Familizing Lonsdale and Le Bras

Having fleshed out the original 5000 headwords of *Lonsdale and Le Bras* as complete lemmas for RQ1, we next familized these for RQ2. We started with cases where the base and derived forms were both contained in the original 5000 lemmas. This was done mainly by alphabetizing the lemmas and assembling nearby members as families. However, we were aware there are many more derived forms that are not frequent enough in themselves to have been included in the first 5000 lemmas. Fortunately, we were able to obtain from Lonsdale the complete headword list from the *Frequency Dictionary* corpus, of which he and Le Bras had used only the first 5000 for the dictionary itself. We fleshed out the additional 20,000 headwords as full lemmas as described earlier and then, largely by hand, integrated all possible further derived forms into the basic 3000 families. For example, "annoncer [to announce]" and its inflections are within the first 3000 families; by searching the larger 25,000 list with hypothetical word roots like "annonce–" and "annonc–" as "starts with" and "contains" terms we discovered "annonceur [announcer]" at the 9000 level and moved it into place as a member of "annoncer" at the 3000 level.

Not every derived word form that was found in either the first 3000 or the the extra 25,000 lemmas was included in our candidate list. We developed a set of transparency criteria to decide whether a derived form could be included in a family:

1. The derived word's meaning had to be clearly related to other members of the proposed family.
2. The derived word's affix(es) had to involve only those normally found in French grammar books.

3. Only one phoneme, or two phonemes in base words longer than eight letters, could be added, deleted, or changed from the stem of the base word in order to be included (eight because enough of the base word is left intact for a learner to have a chance of recognizing it). For example, the following were included:
   a. *angle* [angle] (5 letters) → *ang**u**laire* [angular] with one change,
   b. *approuver* [to approve] (8) → *appro**ob**ation* [approval] with two changes,
   c. *coupable* [guilty] (8) → *c**ul**pabilité* [guilt] with two changes;
   whereas the following were excluded:
   a. *étoile* (6) [star] → *st**ell**aire* [stellar] with at least three changes,
   b. *doute* (5) [doubt] → *du**bi**tatif* [dubious] with two changes,
   c. *nuire* (5) [to harm]→ *n**oc**if* [harmful] with two changes,
4. The relationship between family members had to be plausibly transparent to intermediate French learners, as judged by two out of the three experienced teacher trainers.

Following these procedures, we added close to 19,000 derived words to Lonsdale's 5000 lemmas to form 3000 families of 55,896 word types. In comparison, BNC/COCA's most frequent 3000 families are 19,062 word types.

Nuclearizing Lonsdale and Le Bras

The Lonsdale list when familized clearly needed reduction, which we believed would be best achieved through nuclearization. The first step in nuclearization is to find or assemble a "cross-corpus," against which retention or elimination of each word in the list can be judged. The cross-corpus should reflect or ideally consist of the materials planned for use with a particular group or type of learners or the examination texts used to test them. Or it might reflect a particular student group's subject matter interest (e.g., biology texts) or their current reading level (e.g., graded stories), but for demonstration purposes our cross-corpus is a balanced collection aiming to represent "general French," which resembles the test corpora developed previously. It consists of 1.5 million words, some from different sections or subcorpora of previously sampled corpora (academic writing from the *Chambers–Le Baron* academic corpus), modern adult fiction (Camus other than *L'Étranger*), and children's fiction (another series of *Le Petit Nicolas* stories), and some from novel sources (classic literature from *Frantext*; news stories from the Chambers & Rostand, 2003, corpus of French journalism, and an informal speech collection from Beeching, 1997). The corpus components are roughly equal in size, and the corpus' overall size is modest to assure our software can run over the Internet anywhere French is taught.

The next step in nuclearization is to eliminate words in the complete familized list that do not appear in the cross-corpus at all, the number of which we anticipated with some curiosity. For this we adapted Cobb and Laufer's computer program *Nuclear List Builder* (https://lextutor.ca/freq/nuclear) to run the complete familized lists against our cross-corpus. This software takes as input a family list and returns the same list showing the frequency of every item in the cross-corpus, in both raw numbers and as a percentage of their families, with an option to eliminate any below a certain percentage. In this case, since our goal was to see if size could be reduced without loss of coverage, we removed from the complete family list all words with 0% occurrence in their families. This list we named LFNF-0/3000 (*Liste de fréquence nucléaire française* [French nuclear frequency list] of 3000 families, with items eliminated that are 0% of their families). LFNF-0/3000 consists of 25,152 individual word types, about 30,000 fewer than the original list of 55,896 word types. In other words, fully 30,000 of the words in the original family list do not appear in the 1.5 million word cross-corpus at all. Thus the complete family list can be strongly reduced through nuclearization, but does coverage remain the same?

Figure 1A, where each line shows count and family percentage for each member, illustrates the extent of 0 occurrences in the "nation [nation]" family, taken as an example. Figure 1B shows the same list with no-shows eliminated and statistics removed ready for practitioners. Figure 1C shows LFNF-5

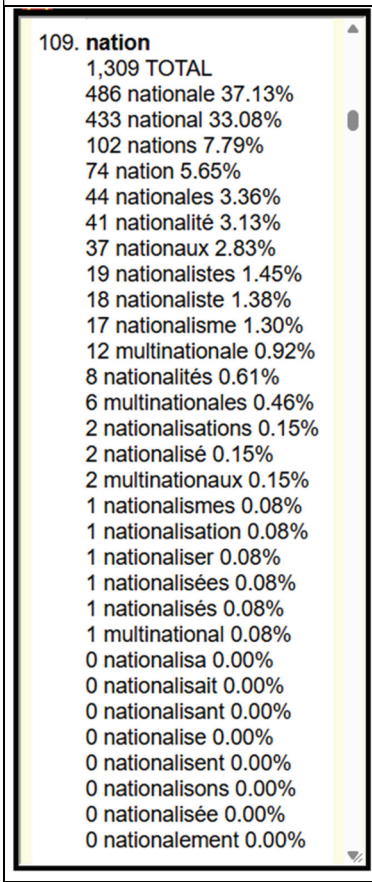
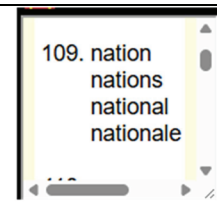
Figure 1A Complete LFNF-0 with totals and percentages in families (sorted by frequency)	Figure 1B LFNF-0 with words of 0.0% occurrences eliminated (sorted alphabetically)	Figure 1C LFNF-5, cut at > 5% (sorted alphabetically)
		

FIGURE 1 Family lists for "nation [nation]." LFNF-0, *Liste de fréquence nucléaire française* [French nuclear frequency list], with items eliminated that are 0% of their families. [Color figure can be viewed at wileyonlinelibrary.com]

with "nation" nuclearized to 5%. Not shown is that there is a rarely applied secondary cutoff of 100 occurrences, to guard against important words being lost from very large families using a percentage cutoff ("suis [am]," is an important word despite its 643 occurrences comprising only 1.8% of its 41-member family, "être [to be]").

RQ2: RESULTS

RQ2 asks how much a list can be reduced without unacceptable loss of coverage. Coverage will be tested in two ways, in the cross-corpus itself and in a novel corpus.

TABLE 2 Coverage tests of LFNF-0.

Text	Size of text (tokens)	LFNF-0 coverage (%)
L'Étranger (novel)	29,729	96.9
Le Petit Nicolas (children's stories)	23,000	97.4
Économie, Chambers academic	90,000	95.9
Philosophie, Chambers academic	40,000	94.7
Dépêche du Midi, Chambers news	118,000	94.6
Parisian Speech	488,000	98.3
Mean (*SD*)		96.24 (1.7)

Abbreviation: LFNF-0, Liste de fréquence nucléaire française [French nuclear frequency list], with items eliminated that are 0% of their families.

LFNF-0/3000 had been reduced from 55,896 word types, with 93% coverage in the cross-corpus, to just over 25,000 word types with 92.53%, almost the same. The answer to RQ2 is, therefore: Yes, the complete familized *Lonsdale* can indeed be cut almost in half with little loss in coverage. This is a testament to the size of the unused and little-used elements of the French word stock.

A stronger coverage test for LFNF-0/3000 is to pitch it, not back against the set of texts it came from (Table 1), but against a similar but novel set of texts. We assembled different test texts with different texts by the same authors (Camus, with *L'Étranger*) or different sections of the same corpora, though we could not find a replacement for *Le Petit Nicolas*. Texts or corpus sections and their sizes in word tokens are shown in Table 2, with coverages in the rightmost column.

The 96.9% coverage of *L'Étranger* was a genuine surprise; the data was run several times to check for error or software hallucination. As shown in Table 2, across a very wide range of text types, from spoken language and children's stories to newspapers and academic articles, LFNF-0 maintains remarkably high level of coverage (96.24%) with low variance ($SD = 1.7\%$).

While creating LFNF-0 was not the primary goal of our study, we believe it is in itself a novel contribution to French pedagogy. It is a master list of the in-use lexicon of the high-frequency zone of French. Though too large for most classroom uses at just over 25,000 word types, it has many potential uses in text profiling and should be equivalent in most ways to the first 3000 families/19,000 word types of the BNC/COCA. This master list can be downloaded from https://lextutor.ca/freq/nuclear/LFNF-0.txt or used to profile texts at https://lextutor.ca/vp/comp. In the remainder of the study, LFNF-0/3000 will serve as the master list from which more focused nuclear sublists can be generated.

RQ3: METHOD

While any number of nuclear lists can be extracted from LFNF-0, by changing the cross-corpus or the cutoff percentage, or both, it is likely that the first 10 cutoffs will provide the most useful lists, with family members removed that are between 1% and 10% of their families. A reduction of 1% will reduce list size a little, while 10% will reduce it significantly, with each reduction involving a loss of coverage. We seek in this part of the study to know, for this particular cross-corpus, the optimal nuclearizing percentage, the size and character of the smallest list with the biggest coverage. Specifically, RQ3 asks how much LFNF-0 can be reduced while still providing 90% coverage in the type of texts learners are likely to be reading.

Generating and comparing lists

Nuclear lists were generated in the manner described for 10 conditions from 1% through 10%, then each was assessed for coverage against three novel corpora chosen to represent three broad learning

TABLE 3 Coverages at different list sizes for three kinds of French.

Percent reduction	Size (word types)	Size reduction	Coverage by text type					
			Literature %	Speech %	Academic %	Mean	*SD*	Mean loss (%)
0	25,152	–	97.83	96.12	92.53	95.49	2.71	–
1	18,473	6686	95.85	94.86	90.9	93.87	2.62	1.62
2	15,127	3346	93.74	93.49	89.66	92.30	2.29	1.50
3	13,168	1959	92.42	92.55	88.54	91.17	2.28	1.13
4	11,576	1592	91.01	91.63	87.52	90.05	2.22	1.12
5	10,458	1118	90.19	90.95	86.63	89.26	2.31	0.79
6	9546	912	87.97	89.73	85.61	87.77	2.07	1.49
7	8831	715	87.20	89.22	84.85	87.09	2.19	0.68
8	8189	642	85.34	88.02	83.49	85.62	2.28	1.47
9	7648	541	83.71	87.15	82.72	84.53	2.33	1.09
10	7143	505	82.97	86.39	81.65	83.67	2.45	0.86
Mean	12,301	1802	88.84	90.92	86.74			
SD	5483.60	1925	4.89	3.16	3.47			

objectives that FLE learners might typically have—to read and discuss news and literature, to read and discuss academic texts, or to participate in informal conversations. The test corpora were three novel extracts from corpora used in earlier parts of this paper: Chambers and Rostand's (2003) journalism corpus, *La Dépêche du Midi* section; Chambers and Le Baron's (2006) academic corpus, social sciences section; and two new segments of 50,000 words each of the Paris (CFPP, Branca-Rosoff, 2012) and Quebec (CFPQ, Dostie, 2015) corpora of unscripted conversational speech.

RQ3: RESULTS

The word type counts of the 10 nuclear lists range from 25,152 in the 0% condition down to 7143 in the 10% condition (Table 3, second column). The main inflection points in list size are from 0% to 1% and then from 1% to 2% (third column), suggesting that, just as there were roughly 30,000 unused word types in the original complete families list, there are another almost 7000 that are just 1% of their families, and a further 3000 that are just 2%, amounting to a further pool of 10,000 little-used words.

Mean coverage across the lists ranges from high at LFNF-0 (95.49%) to acceptable at LFNF-4 and LFNF-5 (90.05% and 89.26%) to inadequate from LFNF-6 on (87.77% and less), though with interesting internal variations. LFNF-4 and LFNF-5 are similar, with 90% coverage in literature and speech but a little less in academic writing, with the difference that LFNF-4 is 1000 word types larger. We therefore propose LFNF-5 with 10,458 word types as a generic list for FLE learners. It is the smallest list with acceptable coverage across two of the three domains, over 90% in literature and speech, though dropping to 86.63% in academic, with a mean coverage of 89.26% ($SD = 2.31$). None of the reduced lists are strong on academic, which may suggest that LFNF-1 (90.9% coverage), with more than 18,000 word types, or LFNF-2 (89.66%) with 15,000, are the greatest reductions possible if the goal is to undertake academic study in French.

The answer to RQ3 is therefore that yes, the size of the 3000-family LFNF-0 can indeed be reduced, from 25,159 words with 95.49% coverage to 10,458 words with 89.26% coverage at LFNF-5, a coverage loss of 7.06% in return for a reduction of 14,701 words, or 57.44%, fewer words to learn. In

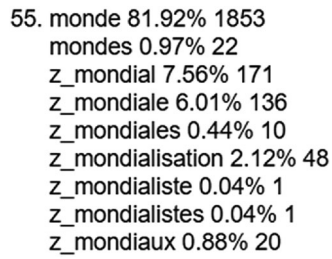


FIGURE 2 Automatic derivations extraction.

cost–benefit terms, this is a good trade. The success of such a nuclearization depends, however, on the assumption that learners already know or can easily acquire the numerous derivational affixes in LFNF-5 that may be new to them. How numerous?

RQ4: METHOD

The matter of MA was raised in the literature review in the context of familizing lemma lists, as this entails a strong expansion in the number of derived words. We saw that while French has a large stock of derivational affixes, many of them are present in the L1's of many FLE learners and may not require much deliberate instruction. Also, in view of previous findings showing large proportions of unused and little-used components of the lexicon of French, we now ask whether this pattern extends to derivational affixes. RQ4 asks how many different derivational affixes a learner has to know and whether learning them imposes an unreasonable burden.

Affix counting

The methodology for this part of the study involved extraction, assembling, counting, and calculating as a percent of the total of affixes in our paradigm list LFNF-5, a manageable task given the reduced size of the list. Here again we are expecting to find a nuclearization or reduction effect. In view of the number of unused and little used inflected and derived word types found in RQ2, we expect to find here also a small number of affixes doing most of the work.

There is no lack of information available about French affixes, just access to a usable pedagogical summary. Mailhot et al. (2020) have compiled a morphological database of 38,840 French derived words, named MorphoLex-Fr, sorted by occurrence in the 50-million-word Subtlex Corpus of film subtitles from which it comes.

To extract the derived forms from LFNF-5, we used List Builder's derivation identifier, which was originally incorporated in the software to help teachers prepare form-focused work with derivations as needed. Each derived form is marked with a preceding "z_" (as shown in Figure 2) to facilitate sorting and extracting a sublist of derived words for a particular nuclear list. The percentages and counts that follow each word refer to the word type's percentage occurrence in each family and then the number of occurrences in the cross-corpus. When the complete list is sorted, it is straightforward to extract the "z_" words and group words by affix.

As a preview of expected results, we can offer Mailhot et al.'s (2020) *Morpholex-Fr* and New et al.'s (2004) *Lexique 3* top 10 affixes (suffixes only, for some reason, and in slightly different parsings): *Morpholex-Fr* ("–eur," "–ion," "–age," "–ment," "–able," "–iste," "–é," "–able," "–ité," and "–if"), and *Lexique* ("–ion," "–ment," "–eur," "–age," "–é," "–able," "–iste," and "–ité"). To be noted is that

corpus coverages are not provided for these affixes whereas in our count they are, at least implicitly, inasmuch as they were all attached to words greater than 5% of their families as found in a corpus.

RQ4: RESULTS

The fruits of our affix count are shown in Table 4, which shows each affix, followed by an example, the variants and inflections the affix was found in, the total occurrences, the percentage of the total (by which the table is sorted), and the cumulative percentage. Our top 10 are similar to *Lexique* and *Morpholex-Fr*'s top 10 except in a slightly different order (though these summaries are not entirely comparable in that we have grouped clearly related affixes ("-tion," "-tions") while the other two approaches differentiate them by gender, number, and possibly other distinctions).

An interesting if predictable finding from our count is that 2617 derived words are present in the list of 10,458 words, or about 25%. The more interesting finding to jump out of the Table 4 is that despite the large stock of French affixes, just five of them ("–tion," "–ment," "re–," "–ité," and "–eur") comprise more than 50% of the number in use, at least in this particular cross-corpus, which incidentally are all cognate with both English and the Romance languages. Just 20 account for 94%. In comparison, 20 account for 80% in the English NFL-7/3000. Thus, French once again shows a similar but greater nuclear effect than English, reflecting the extent of its unused and little-used holdings. It would be interesting to compare the affix list of Table 4 to the contents of a comprehensive textbook for learning derivational morphology in FLE if one could be found.

We conclude that with 50% of derivational affixes directly transferable from L1 for many FLE learners, though admittedly with small differences, extensive morphology training is probably not required for receptive purposes. English FLE learners who are taught "soigner [to care for]," can be assumed able to interpret "soigneur [care giver]" when they see it, without extra instruction, from their knowledge of "–er" (other than perhaps a one-time heads-up that "–eur" is like *–er*, or for a Spanish learner that it is like "–or"). Or in cases of non-English-like employments ("hauteur [height]," and "douceur [sweetness]") to have at least an initial hypothesis about a plausible partition of morphemes. So the answer to RQ4 is that no, adding derived word forms to a pedagogical word list poses a small, not unreasonable, learning burden for most learners.

For the others, or if for any reason a pedagogy of MA for derivations was desired, the most frequent 20 affixes, as now identified, could provide the core of it. For example, if a teacher emphasized—whether through worksheets, profiled texts, or ad hoc interventions—the first three affixations in Term 1, the first 10 in Term 2, and the first 20 in Term 3, then 94% of the total would be well known and abundantly practised, and we venture the remaining 6% would be learned independently. Or if a more explicit MA pedagogy was needed or desired, we offer an approach to it in the next section.

DISCUSSION

The value of list nuclearization validated by Cobb and Laufer (2021) for English has proven transferable to French. Indeed, a reduction process that was powerful for English is more powerful for French, because of the vast proportion of French words and affixes that are unused and little used. Some form of reduction is clearly needed to identify the in-use core of the language, rather than leaving teachers and learners to find it for themselves. For English, nuclearization shortened and tightened a good list, a matter of degree; for French, nuclearization makes a pedagogical list possible, a matter of existence. The complete BNC/COCA list of 3000 families was reduced from 19,062 to 7293 types (38% of the original); the complete familized Lonsdale and Le Bras list of 3000 families was dramatically reduced from 55,896 to 10,458 words (18.7% of the original, almost exactly half).

TABLE 4 Proportion of affixes in the 2617 derived forms of LFNF-5 (10,458 words).

	Base affix	Example	Frequency by inflection/further derivation	Total	Percentage of total	Cumulative percentage
1	–tion	abolition	–tion 98; –tions 51; –(a|i)tion 281; –(a|i)tions 151; –sion 119; –sions 36	736	25.1	25.1
2	–ment	acharnement	–ment 346; –ments 34	380	13.0	38.1
3	re–	rebaptisé	re– 100; ré– 65	165	5.6	43.7
4	–ité	absurdité	–ité 123; –ités 22; –eté 12; –etés 1; –ïté 1	159	5.4	49.1
5	–eur	procureur	–eur 84; –eurs 64; –rice 4; –rices 2	154	5.3	54.4
6	–ique	téléphonique	–ique 66; –iques 38; –iquement 16	120	4.1	58.5
7	–al	mondial	–al 39; –ale 40; –als 1; –ales 11; –aux 12; –auté 5	108	3.7	62.2
8	–ant	passant	–ant 30; –ants 12; –ante 8; –antes 3; –ance 42; –ances 8	103	3.5	65.7
9	in–	inconvénient	in– 100	100	3.4	69.1
10	–eux	affreux	–eux 43; –euse 24; –euses 6; –eusement 17	90	3.1	72.2
11	–able	comparable	–able 44; –ables 18; –ible 12; –ibles 6; –ablement 6	86	2.9	75.1
12	–el	habituel	–el 29; –els 6; –elle 17; –elles 9; –ellement 20	81	2.8	77.9
13	–ier	fermier	–ier 35; –iers 17; –ière 16; –ières 9	77	2.6	80.5
14	–aire	universitaire	–aire 45; –aires 31	76	2.6	83.1
15	–age	personnage	–age 51; –ages 12	63	2.2	85.3
16	–isme	racisme	–iste 21; –istes 17; –isme 23	61	2.1	87.4
17	–if	sportif	–if 6; –ifs 9; –ive 21; –ives 8; –ivement 14	58	2.0	89.3
18	dé–	découvrir	dé– 50	50	1.7	91.0
19	–ien	Parisien	–ien 22; –ienne 8; –iens 18; –iennes 1	49	1.7	92.7
20	–ain	Mexicain	–ain 10; –aine 15; –ains 8; –aines 7	40	1.4	94.1
21	–ure	blessure	–ure 19; –ures 7	26	0.9	95.0

(Continues)

TABLE 4 (Continued)

	Base affix	Example	Frequency by inflection/further derivation	Total	Percentage of total	Cumulative percentage
22	–oir	transitoire	–oir 5; –oirs 1; –oire 12; –oires 6; –oirement 2	26	0.9	95.9
23	–esse	délicatesse	–esse 20; –esses 5	25	0.9	96.7
24	–at	artisanat	–at 14; –ate 3; –ats 4; –ates 3	24	0.8	97.5
25	–ence	existence	–ence 6; –ences 8	14	0.5	98.0
26	–ais	japonais	–ais 8; –aise 6; –aises 0	14	0.5	98.5
27	–ade	fusillade	–ade 6; –ades 5	11	0.4	98.9
28	–ois	Danois	–ois 7; –oise 3; –oises 1	11	0.4	99.2
29	–ième	deuxième	–ième 9; –ièmes 0	9	0.3	99.6
30	–éen	Européen	–éen 1; –éens 1; –éenne 1; –éennes 1	4	0.1	99.7
31	–esque	gigantesque	–esque 2; –esques 0	2	0.1	99.8
32	–o	franco	–o 8	6	0.2	100.0
Totals				2928	100.0	

Note: It can be observed there are more derivational affixes (2928) than derived word types (2617). That is because many words in the list have more than one affix (e.g., "rémarquablement" with "–able" and "–ment").

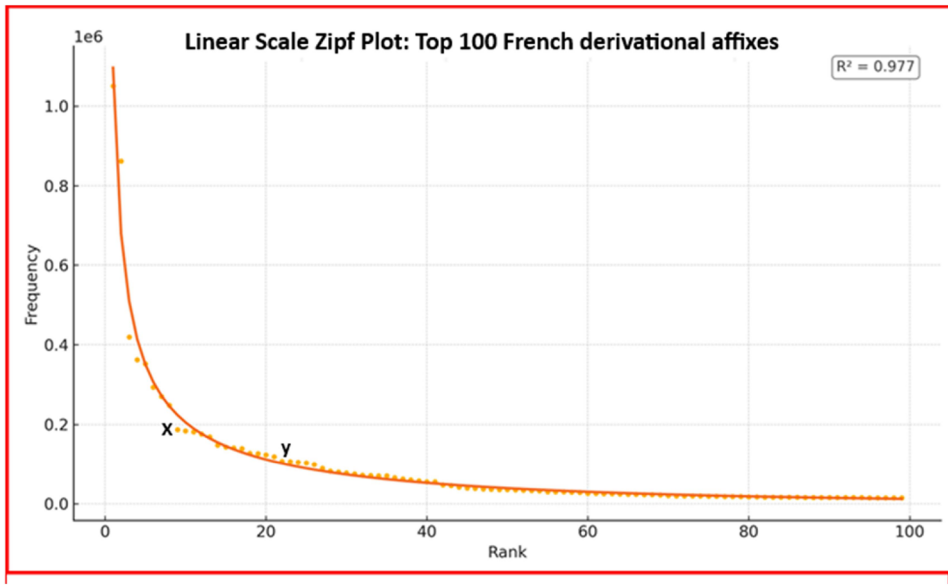


FIGURE 3 Derivational affix distribution for LFNF-5/3000 in Fr_combo.txt (2.7 million words). LFNF-5, Liste de fréquence nucléaire française [French nuclear frequency list]. *Note*: Rendered by OpenAI on 6 June 2025, from data points in Table 4, Columns 2 and 5. The points "x" and "y" approximate a zone of mid-frequency. [Color figure can be viewed at wileyonlinelibrary.com]

CEFR alignment

It has been noted that the CEFR reform (Trim, 2012) was not accompanied by a set of word lists for reasons that we have shown are no longer pertinent. One reason was that the spoken emphasis of CEFR necessitated speech corpora and lists which were not available at the time; now there are many of these for different languages, some of which we have incorporated throughout this paper. Another was that in the early CEFR period many of the target languages had no corpora to develop lists from; this is no longer the case for any of the target languages. Finally, the problem of one list becoming dominant can be avoided by generating different nuclearizations of the same list for different purposes and situations, which can happen "close to the point of learning" (p. 25).

We propose that a nuclearization of LFNF-0 based on a cross-corpus of CEFR materials could produce a more useful set of CEFR word lists than was achieved by the Kelly or CEFRLex approaches. As a proof of concept, we have obtained a 7000 word mini-corpus of six graded 'Maigret' stories described by their publisher (CLE international) as CEFR level A2 to C2, or lower intermediate to advanced, and run them against our LFNF-5. The two stories described as A2 achieved 90% coverage with just the first 512 families of LFNF-5; the four described as B2–C1 achieved 90% employing the entire first 1000 families plus 408 from the second 1000 and 235 from the third. In other words, a group of FLE learners reading these stories and working through LFNF-5 at the same time, whether as a list or as texts profiled to it, would meet the same words in both places and enjoy a large degree of word recycling.

A bigger picture

Once a nucleus to the French lexicon has been identified, it takes on an air of inevitability. It is after all but an instance of Zipf's Law, by which any sizeable distribution of language features will configure

itself with a few used much and many used little. Figure 3 shows the Zipfian distribution of the most frequent 100 affixes from Table 4, chosen because it is a convenient number of items to show on a page (though we have found a similar distribution for every other dimension of LFNF-5).

In this case, the data points follow a Zipfian model particularly closely, with 97.7% of their variance explained by the fitted curve. So it is a strong Zipfian distribution. This confirms that nuclearization has not produced a word list with a skewed or irregular distribution, as would have been the case if too small a cross-corpus had been used, or one with just one type of text, or one that was out of balance mixing, say, children's stories and academic monographs and nothing in between.

What is important for language learning about a strong Zipfian curve? It is that the frequency descends "gracefully" such that the middle frequency is populated. And as we saw in the case of the French textbooks, it is the middle frequencies that are typically missing from vocabulary development in instructed French. Roughly 10 affixes on the left of the figure occur very frequently, while many others trail off to lower right and occur rarely. It is the ones in the middle that are interesting, roughly items 10 to 20 in Table 4 (marked "x" and "y" in the figure—namely, "–eux/–euse," "–able/–ible," "–el," "–ier/–ière," "–aire," "–age," "–isme," "–if/–ive," "dé–," "–ien," and "–ain"). This is the elusive middle that native speakers know but may not be aware of (McCrostie, 2007) and that a Zipfian representation can identify and bring into play.

Zipfian distributions are key to usage-based approaches to language acquisition such as that developed by Ellis et al. (2016). The guiding question of researchers in this approach is: How is language acquired if there is no (Chomskyan) language acquisition device? The answer proposed is that it is acquired through normal cognitive processes operating on a small set of high-frequency prototypes at all levels of language, like the handful of affixes in the upper-left part of Figure 3. Thus, through abundant exposure, learners notice, practice, normalize, chunk, apply, and transfer the most useful parts of the language first and build a scaffold to handle the remainder at the bottom right of the graph.

Abundant exposure to the prototypical nuclei of a language is gifted to children in their L1s, through modified input and children's inherent ability to ignore language outside their zones of development. But how do older second or foreign language learners identify and employ exemplars and prototypes? If they do, it could be because a teacher possessed and shared accurate intuitions about linguistic frequency, or else employed the resources of a frequency list drawn from an appropriate corpus. The recurring argument throughout this article is that, absent exceptional intuition, a small, high coverage frequency list can identify first the prototypical nuclei of a language and thereafter the elusive mid-frequency zones that facilitate lexical growth beyond "between 800 and 1,000 words" (Milton, 2006, p. 193).

To be noted is that usage or input-driven learning is not very *didactique* in nature. The role of the teacher and curriculum are mainly to provide the motivation to communicate and the input that learners will largely put together for themselves through interaction with peers and knowledgeable others. Not that crafting optimal input is simple in a foreign language—as we hope to have shown.

We believe our nuclearization process and its accompanying software allow practitioners a means to exploit these principles on behalf of their learners. Large frequency lists like BNC/COCA or LFNF-0 contain these prototypical zones, just well hidden; nuclearization renders them salient and hence learnable. Nuclear lists are essentially distilled frequency lists, frequency lists of frequency lists; they recreate the salient prototypes of L1 acquisition in a second or foreign language. And with a profiler running the appropriate nuclear lists over texts that their learners find motivating, teachers can provide natural input that is naturally sequenced for learning.

Morphological awareness in usage-based learning

Following the resolution of RQ4 we are in a position to return to the question of MA with regard to derivations, an important issue for our reviewers in view of our plan to incorporate derived words

within nuclear families. Does a usage-based approach have a place for MA or does it favor learning from usage alone?

Admittedly, usage-based or input-driven learning does not have an obvious place for didactic or top-down pedagogies like MA training. "Top down" refers to the practice of teaching rules prior to the instances they apply to (e.g., "Here is what '–ment' means," when learners have not yet met or noticed a commonality in "gentiment [nicely]" and "heureusement [luckily]." Once identified and explained, the affix thus stands ready for application to new derived words as they enter the system, and there is less risk that "drôle [funny]" and "drôlement [amusingly]" may forever remain unlinked items (as Schmitt & Zimmerman, 2002, show is a common result in even advanced uninstructed learning). An issue of possible relevance is that the massive input of L1 acquisition is rarely replicated in L2, which may leave a greater role for didactic approaches. FLE learners' opportunities for abstracting morphological patterning on a strict usage basis are likely to be more limited.

In fact, input theorists, especially when discussing EFL and FLE learning environments, indeed endorse lightly pedagogicalizing emergent processes through increasing salience, highlighting analogies, and facilitating noticing. Ellis (1997) did not deny there is "any role of pedagogical rules in language learning" or dispute that "some parts of the [language acquisition] environment can be made more salient" (e.g., "by 'grammatical consciousness raising' or 'input enhancement' or 'focus on form'"(p. 21). An example of input enhancement would be highlighting different affixes or affix types throughout a text with different colors. From such hints of a usage and input pedagogy, data-driven learning (DDL) has arisen as a branch of computer-assisted language learning (CALL) dedicated to rendering language patterns salient through computation.

Data-driven morphological awareness

For a specific example of a DDL version of MA we turn back to Thorndike (1941) for inspiration, who generated the first pedagogical list of 90 English derivational suffixes along with usage-based approaches to learning them. For each affix, he proposed learners "spend five minutes looking at a list of words made with it" (Thorndike, 1941, p. 59). This disarmingly simple idea is rudimentary input-based learning. It is also somewhat inefficient—Are all 90 suffixes of equal importance? How long does it take to make the lists? How do learners know when they have isolated the affix or inferred the correct meaning?—but is primed for a DDL enhancement. A computer program linked to a range of corpora (such as *Concordancier français* [French Concordancer] at https://www.lextutor.ca/conc/fr/) can easily produce a list of sentences bearing a particular affix. The brief of DDL is to simulate input/usage-based learning processes more efficiently or saliently than they appear in nature in the timeframe of L2 acquisition.

DDL software can corpus-generate a Thorndike list of derived words of any size, highlight their affixes and immediate syntactic environments, access the texts each originated in, and ask learners to integrate this information and transfer it to new words. For a simple example, Figure 4 shows Part 1 of the process, examples of software-generated input salience enhancers with random "–ment" and "–tion" words taken from an informal corpus of *Le Monde* news articles from 1986 provided by Da Silva (personal communication, 2000), but any number can be similarly provided and the difficulty level raised or lowered by choosing other corpora, or by including cases where "–ment" is "not used in an English way" ("déclenchement [unclenching]," "abaissement [lowering]," i.e., where "–ment" is not equivalent to "–ly"), or where affixation has involved a substantial change to the base word ("culpabilité [guilt], from "coupable" [guilty]". Part 2 of the process (Figure 5) is to apply the knowledge gained in Part 1 to a fresh set of words from the same or a comparable corpus. Developing a data-driven syllabus of the most frequent 30 affixes whether on screen or paper would be a useful instructional design project.

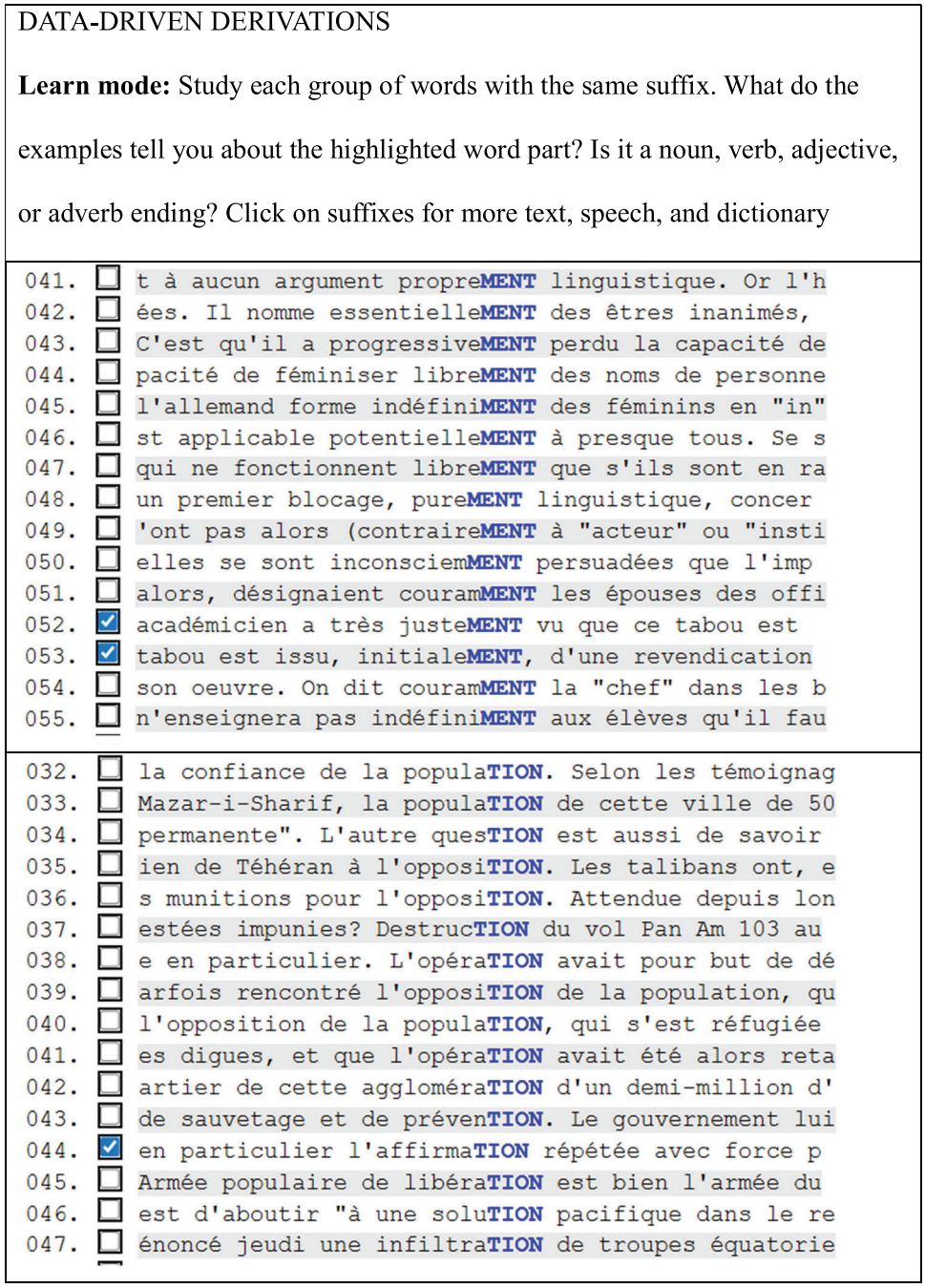


FIGURE 4 Input-oriented morphology trainer. *Note*. Checked boxes indicate that learner has accessed supplementary information. [Color figure can be viewed at wileyonlinelibrary.com]

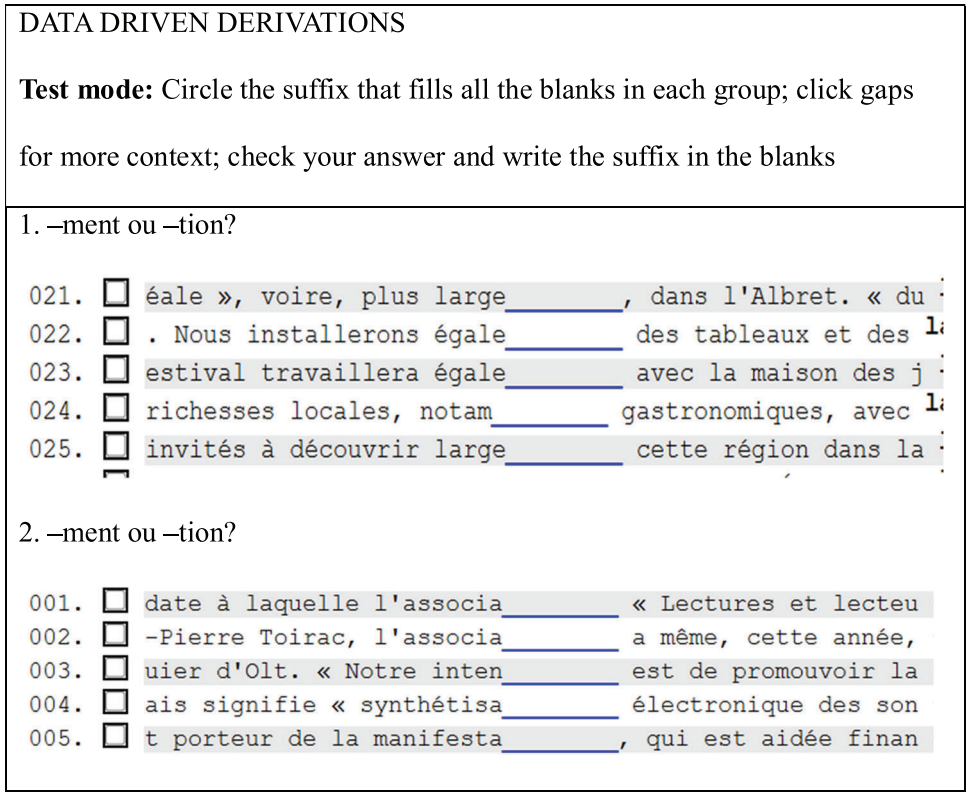


FIGURE 5 Follow-up morphology tester. Technical instructions for producing these or similar materials on- or off-line, group or individual instruction appears in the Online Supporting Information. A good collection of morphological awareness training ideas including Thorndike's is found in Chapter 9, "Word Parts" of Nation's (2013) book, many of them adaptable to both French and to data-driven learning treatments. [Color figure can be viewed at wileyonlinelibrary.com]

CONCLUSION

In this article, we have presented a new word list for French: the LFNF. It is the first pedagogical list for French that is based on word families and small enough to be readily used. Though containing relatively few word families and family members, it provides a high level of coverage across a range of texts that French learners typically encounter in their reading. Different versions of the basic list are available at different grain sizes, making it useful for learners with different goals and proficiency levels (an idea detailed in the Online Supporting Information). As we saw, the same 3000 families can provide a 10,000 word list for literary texts at 5% reduction, or a 20,000 word list for academic texts at 1%. Either is "small" compared to the alternatives. Future work with this list will include testing its claims and usability, and further tweaking its entries, for which we depend on the helpful critiques of journal readers, research colleagues, and teachers and learners of French around the world.

AUTHOR CONTRIBUTIONS

Thomas Cobb: Conceptualization (supporting); methodology (equal); formal analysis (equal); data curation (supporting); investigation (equal); resources (lead); software (lead); writing—original draft (lead); writing—review and editing (lead); and validation (supporting). **Christina Lindqvist**: Conceptualization (lead); methodology (equal); formal analysis (equal); data curation (lead); investigation (equal); resources (supporting); software (supporting); writing—original draft (supporting); writing—review and editing (supporting); and validation (lead). **Mårten Ramnäs**: Conceptualization (lead);

methodology (equal); formal analysis (equal); data curation (lead); investigation (equal); resources (supporting); software (supporting); writing—original draft (supporting); writing—review and editing (supporting); and validation (lead).

In the case of our study, **Conceptualization** includes (in addition to its taxonomic meaning) needs analysis; **Data curation** includes ongoing surveillance of the accuracy and appropriacy of French language usage; **Investigation** includes review of literature; **Resources** includes corpus building for the numerous informal text corpora employed in this study; and **Validation** includes steps toward the application of our findings to the broader world of French language education.

DATA AVAILABILITY STATEMENT

All data references are accompanied by URLs where it may be obtained.

ORCID

Thomas Cobb https://orcid.org/0000-0003-4287-8789
Christina Lindqvist https://orcid.org/0000-0003-1590-1437
Mårten Ramnäs https://orcid.org/0000-0002-2531-0019

REFERENCES

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Language Journal*, *91*, 659–663. http://www.jstor.org/stable/4626093

Antes, T. (2023). A general service list for French? Teaching the vocabulary that matters, *Language Teaching*, *56*, 570–573. https://doi.org/10.1017/S0261444823000289

Anthony, L. (2024). List of BNC written lemmas. https://www.laurenceanthony.net/resources/wordlists/bnc_wordlist.zip

Arndt, R. (2025). French vocabulary in the *T'es branché* series: A corpus study. *The Language Learning Journal*, 1–13. https://doi.org/10.1080/09571736.2025.2453526

Baudot, J. (1993). *Fréquence d'utilisation des mots en français écrit contemporain* [Frequency of word use in contemporary written French]. Les Presses de l'Université de Montréal.

Beeching, K. (1997). The case for spoken corpora. *Applied Linguistics*, *18*, 374–394. https://doi.org/10.1093/applin/18.3.374

Bell, C., & McLachlan, A. (2012). *Studio 3 vert Pupil Book* (11–14 French). Pearson Education.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2012). *Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*. http://cfpp2000.univ-paris3.fr/CFPP2000.pdf [Corpus download http://cfpp2000.univ-paris3.fr/]

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*, 1–22. https://doi.org/10.1093/applin/amt018

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Camus, A. (1942). *L'Étranger*. Gallimard.

Carrère, E. (2000). *L'adversaire*. Gallimard.

Chambers, A., & Le Baron, F. (2006). *Chambers-Le Baron corpus of research articles in French*. https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2527

Chambers, A., & Rostand, S. (2003). *The Chambers-Rostand corpus of journalistic French*. Oxford Text Archive. https://llds.phon.ox.ac.uk/llds/xmlui/handle/20.500.14106/2491

CLE international. L'éditeur du français langue étrangère [The publisher for FLE]. https://cle-international.com/adolescents

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development*, *47*, 15–33. https://link.springer.com/article/10.1007/BF02299631

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*, 834–871. https://doi.org/10.1111/lang.12452

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. https://www.coe.int/lang-cefr

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge. https://doi.org/10.4324/9780429291586

Dang, T. N. Y., & Webb, S. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*, 1–25. https://doi.org/10.1177/1362168820911189

Davau, M., Cohen, M., & Lallemand, M. (1972). *Dictionnaire du français vivant* [*Dictionary of living French*]. Bordas.

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal*, *36*, 167–180. https://doi.org/10.1080/09571730802389991

Davies, M. (2008). Corpus of contemporary American English. https://www.english-corpora.org/coca/
De la Maya Retamar, G., & Mora Ramos, I. (2019). Estudio del conocimiento léxico en FLE de estudiantes españoles de Secundaria [Study of Spanish secondary FLE students' lexical knowledge]. *Revista Complutense De Educación*, *30*, 527–543. https://doi.org/10.5209/RCED.57773
Department for Education, UK. (2022). *GCSE French, German and Spanish subject content*. https://www.gov.uk/government/publications/gcse-french-german-and-spanish-subject-content
Dostie, G. (2015). *CFPQ (Corpus de français parlé au Québec)*. Université de Sherbrooke, Quebec. https://applis.flsh.usherbrooke.ca/cfpq/
Duncan, L., Casalis, S., & Colé, P. (2009). Early metalinguistic analysis of derivational morphology: Observations from a comparison of English and French. *Applied Psycholinguistics*, *20*, 405–440. https://doi.org/10.1017/S0142716409090213
Ellis, N. C. (1997). The epigenesis of language: Acquisition as a sequence learning problem. In A. Wray & A. Ryan (Eds.) *Evolving models of language: Papers from the Annual Meeting of the British Association for Applied Linguistics* (pp. 41–57). Multilingual Matters.
Ellis, N., Römer, U., & O'Donnell, N. (2016). *Usage-based approaches to language acquisition & processing: Cognitive & corpus investigations of construction grammar*. Wiley. https://doi.org/10.1111/lang.12193
Finlayson, N., & Marsden, E. (2024). Wordlists for French, German, and Spanish. https://osf.io/jyshp/
Fejzo, A. (2020). Connaître les rouages de la morphologie française: un levier puissant pour l'apprentissage des mots [Knowing the tricks of French morphology: A powerful tool for learning words]. *Correspondances: La Revue Web Sur La Valorisation Du Français En Milieu Collégial*, *25*(6), 1–17. https://correspo.ccdmd.qc.ca/index.php/document/connaitre-les-rouages-de-la-morphologie-francaise-un-levier-puissant-pour-lapprentissage-des-mots/
François, T., Gala, N., Watrin, P., & Fairon, C. (2014, 26–31 May). FLELex: A graded lexical resource for French foreign learners [Paper presentation]. The *9th International Conference on Language Resources and Evaluation* (LREC 2014), Reykjavik, Iceland.
Gala, N., & Rey, V. (2008). POLYMOTS: Une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques [A database of derivational constructions in French based on phonological roots]. TALN Conference, Avignon, France. https://polymots.huma-num.fr/
Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. https://doi.org/10.1093/applin/amt015
Goscinny, R. (1960). *Le Petit Nicolas*. Editions Denoël.
Gougenheim, G., Rivenc, P., Sauvageot, A., & Michéa, R. (1956). *L'élaboration du français fondamental, 1er et 2e degrés* [The development of basic French word lists, Levels 1 & 2]. Didier.
Iwaizumi, E., & Webb, S. (2022). Measuring L1 and L2 productive derivational knowledge: How many derivatives can L1 and L2 learners with differing vocabulary levels produce? *TESOL Quarterly*, *56*, 100–129. https://doi.org/10.1002/tesq.3035
Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL* (pp. 125–127).
Kieffer, M., & Lesaux, N. (2007). Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. *The Reading Teacher*, *61*, 134–144. https://doi.org/10.1598/RT.61.2.3
Kokkinakis, S., & Volodina, E. (2013). Corpus-based approaches for the creation of a frequency based vocabulary list in the EU Project KELLY—Issues on reliability, validity, and coverage. *US-China Foreign Language*, *11*, 211–226. https://www.davidpublisher.com/Public/uploads/Contribute/552f78b55d048.pdf
Kučera, H., & Francis, W. (1967). Computational analysis of present-day American English. *Providence*, RI: Brown University Press.
Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*, 1076–1085. https://doi.org/10.1002/tesq.3004
Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learner's vocabulary size and reading comprehension. *Reading in a Foreign Language 22*, 15–30. https://doi.org/10.64152/10125/66648
Lindqvist, C. (2018). Le développement de la taille du vocabulaire en français L2 chez les élèves suédophones [The development of vocabulary size in Swedish speaking French as a second language students]. *Synergies Pays Scandinaves*, *11/12*, 151–161. https://gerflint.fr/Base/Paysscandinaves11_12/lindqvist.pdf
Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners* [CD-ROM]. Routledge. https://doi.org/10.4324/9780203883044
Mailhot, H., Wilson, M. A., Macoir, J., Deacon, S. H., & Sánchez-Gutiérrez, C. (2020). MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, *52*, 1008–1025. https://doi.org/10.3758/s13428-019-01297-z
Marsden, E., Finlayson, N., & Anthony, L. (2023). Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, *118*, 103–122. https://doi.org/10.1016/j.system.2023.103122
Mascie-Taylor, H., & Honnor, S. (2001). *Encore Tricolore: Nouvelle Edition*. Nelson Thornes.
McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *RELC Journal*, *38*, 53–66. https://doi.org/10.1177/0033688206076158
McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit, *Applied Linguistics*, *39*, 823–845. https://doi.org/10.1093/applin/amx003

Milton, J. (2006). Language lite? Learning French vocabulary in school, *Journal of French Language Studies*, *16*, 187–205. https://doi.org/10.1017/S0959269506002420
Milton, J., & Hopwood, O. (2022). Vocabulary loading in the Studio textbook series: A 40% decline in the vocabulary input for French GCSE. *The Language Learning Journal*, *50*, 667–683. https://doi.org/10.1080/09571736.2021.1916572
Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language, *ITL Review of Applied Linguistics*, *107/108*, 17–34. https://doi.org/10.1075/itl.107-108.02mil
Ministère de l'éducation nationale et de la jeunesse [Ministry of National Education and Youth, France]. (2020). Eduscol : Liste de fréquence lexicale. https://eduscol.education.fr/186/liste-de-frequence-lexicale
Modiano, P. (2001). *Le Petit Bijou*. Gallimard.
Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59
Nation, P. (2012). *The BNC/COCA word family lists*. https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf
Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656
Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins. https://doi.org/10.1075/z.208
Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–12. https://openaccess.wgtn.ac.nz/articles/journal_contribution/A_vocabulary_size_test/12552197/1/files/23375153.pdf
New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 3: A new French lexical database, *Behavior Research Methods*, *36*, 516–524. https://doi.org/10.3758/bf03195598
Oxford University Computing Services. (1995). *The British National Corpus, v.2 (BNC World)*. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/
Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018
Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*, 145–171. https://doi.org/10.2307/3588328
Stæhr, L. (2008). Vocabulary size and the skills of listening, reading, and writing. *The Language Learning Journal*, *36*, 139–152. https://doi.org/10.1080/09571730802389975
Theisen, T. (2019). *T'es branché?* EMC Publishing.
Thorndike, E. (1941). *The teaching of English suffixes*. Teacher's College, Columbia University.
Trim, J. (2012). The Common European Framework for Reference for languages and its background: A case study of cultural politics and educational influences. In M. Byram & L. Parmenter (Eds.), *The common European framework of reference: The globalisation of language education policy* (pp. 14–33). Multilingual Matters.
Tschichold, C. (2012). French vocabulary in *Encore Tricolore*: Do learners have a chance? *Language Learning Journal*, *40*, 7–19. https://doi.org/10.1080/09571736.2012.658219
Tutin, A., & Grossmann, F. (2014). L'écrit scientifique : du lexique au discours. *Rennes : Presses de l'Université*.
West, M. (1953). *A general service list of English words*. Longman.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cobb, T., Lindqvist, C., & Ramnäs, M. (2026). A nuclear families word list for French. *Modern Language Journal*, 1–28. https://doi.org/10.1111/modl.70021