2012-12-13

A Corpus-Based Evaluation of the Common European Framework Vocabulary for French Teaching and Learning

Francoise S. Kusseling
Brigham Young University - Provo

A Corpus-Based Evaluation of the Common European Framework Vocabulary for French Teaching and Learning

Françoise Kusseling

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Dee Gardner, Chair
Deryle Lonsdale
Mark Davies
Randall Davies
Richard Sudweeks

Department of Instructional Psychology and Technology

Brigham Young University

December 2012

ABSTRACT

A Corpus-Based Evaluation of the Common European Framework
Vocabulary for French Teaching and Learning

Françoise Kusseling
Department of Instructional Psychology and Technology, BYU
Doctor of Philosophy

The CEFR French profiles have been widely used to teach and evaluate language instruction over the past decade. The profiles were specifications of vocabulary that have been largely untested from a corpus-based, empirical perspective. The purpose of this dissertation was to evaluate the CEFR profiles by comparing their content with two sizable contemporary corpora. This study quantified and described the vocabulary overlap and uniqueness across all three of these resources. Four areas of overlap and three areas of uniqueness were analyzed and identified. Slightly over 40% of the lexical content was common to the three resources studied. Additionally, 16.3% was unique to the CEFR. The remaining CEFR content overlapped with one or the other of the two corpora used for the evaluation. The findings led to the general recommendation of keeping about 60% of the current CEFR content and adding a little over 19,000 vocabulary items to the overhauled CEFR profiles.

Keywords: French, CEFR profiles, corpora, vocabulary, evaluation

ACKNOWLEDGEMENTS

I express great appreciation to each member of my dissertation committee and to my first chair, Dr. Wilfried Decoo, who retired in 2011. I also value the input and suggestions of my editor, Dr. Tricia Stoddard. They all provided expert assistance in research, writing, and technical aspects of the dissertation. I also express sincere thanks to the faculty of the BYU Department of French and Italian, who gave me opportunities to work with them and with learners of French as a foreign language. Last but not least, I thank the faculty and staff of the BYU Department of Instructional Psychology and Technology for their strong support, encouragements, and direction.

I dedicate my work to all those who will learn from it, and to my dear mother, Anita Michèle, deceased 5 July 2011. She wanted me to succeed in this endeavor, and, as a youth, aspired to teach French. She did so beautifully in our home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Since the 1970s the Council of Europe Language Policy Division, now in Strasbourg, France, has mobilized researchers and pedagogues to introduce methodological innovations for language instruction programs and to develop a communicative teaching approach which would facilitate the exchange of people and ideas within the European community and abroad. With this perspective in mind, Van Ek, Trim, and colleagues developed an operational model for teaching basic language skills and common everyday vocabulary people (e.g. tourists, business people, migrants) might need to perform tasks independently using a foreign language (Van Ek, 1975, 1976; Van Ek et al.,1977; Van Ek & Trim, 1984). The English *Threshhold Level* and the French *Un Niveau Seuil*, both vocabulary profiles, were produced as descriptive references of linguistic proficiency levels intended to allow comparability across European languages.

To improve and monitor learners' linguistic autonomy, the Language Policy Division commissioned European researchers to produce vocabulary profiles tiered by proficiency level. These profiles were designed to use a communicative approach based on the Common European Framework of Reference (CEFR) for languages. Nine countries finalized or are currently developing vocabulary Reference Level Descriptions (RLDs): the Czech Republic, Germany, the United Kingdom, France, Georgia, Greece, Italy, Spain, and Portugal. In France, the vocabulary descriptions are commonly known as *Référentiels* and in English as *Profiles*. The term *profiles* used in this text refers to the French RLD developed from the CEFR.

These ongoing European efforts stemmed from a will to evaluate the proficiency of learners of foreign languages and motivate them to communicate, identify tasks they are able to perform in their foreign languages, and self-assess their linguistic skills. The initiative was vital for curricular instructional design, skill development, strategy training, graded reading, and

psychological testing and placement. The production of French vocabulary profiles would also have an impact on technology since new methods of text evaluation and readability resort to word lists derived from frequency studies of mega-corpora to improve formulas and feedback. The results of these studies would, in turn, heighten reading capabilities.

Learners of a foreign language need to acquire a critical mass of vocabulary to reach an advanced level of language proficiency (Coady & Huckin, 1997; Grabe, 1986). It is however difficult to determine how many words and which words best represent that critical mass. Frequency studies are used to establish a list of the most commonly used vocabulary. Based on the results of these studies curriculum developers can then decide which words an individual might need in order to be considered adept at a specific language level (Adolphs, 2006; Adolphs & Schmitt, 2003, 2004; Coxhead, 2000; Schonell, Meddleton, & Shaw, 1956; Zipf, 1935). Results from frequency studies have also been used to calculate readability scores. However, to date no study has been done to test the CEFR French profiles against the rankings of frequency studies in order to improve vocabulary selection and distribution by level. It is not known how well the French profiles represent the vocabulary a learner will likely encounter .

To develop the French tiered profiles, the CEFR French research team (Beacco, Porquier, & Bouquet, 2004; Beacco & Porquier, 2007; Beacco, Lepage, Porquier, & Riba, 2008; Beacco, Blin, Houles, Lepage & Riba, 2011) opted to rely on the RLDs, established knowledge regarding learners and acquisition sequences, common curricular goals, collective experience of teachers and evaluators, and CEFR criteria-related examples of learner productions. Unlike their English colleagues, they did not have at their disposal the corpus linguistics data that would have allowed them to conduct scientific studies and evaluate how the profiles might compare with ranked frequencies calculated from large to mega French electronic corpora.

It was unknown what coverage of the stable French general core the more subjective, notional, task-based profiles provided. However, these vocabulary profiles have been and were to be used for two purposes. The first was to make the teaching and learning of French as a foreign language comparable to the teaching and learning of other European languages and secondly to define levels of vocabulary proficiency.

Focus of the Study

This research dealt with curricular vocabulary input, and, more particularly, with vocabulary selected to teach beginning and intermediate level French as a foreign language. The study investigated what learners of the French language need in the way of vocabulary instruction to move progressively from beginning to advanced proficiency level in order to function effectively in an academic or work environment. It evaluated CEFR vocabulary content and its apportioning by proficiency levels for instructional purposes. More specifically this study looked at criteria governing vocabulary selection and distribution used for French language learning and instruction. It also addressed what threshold should be reached to meet specific language needs to be considered proficient at a specific level.

Determination of priorities in vocabulary syllabus content does not automatically offer a method for teaching words progressively by proficiency level. The challenge is to plan continuous and systematic selection and distribution of content for progressive language instruction from up-to-date inventoried oral and written text. A plan like this implies the idea of gradual levels determined on the basis of adult learners' needs which are assumed to include the acquisition of sufficient vocabulary to function as an independent communicator. Levels of progression, in turn, imply dynamic movement between levels toward the acquisition of a critical mass of vocabulary knowledge. The desired outcome of vocabulary input design would be a

needs-specific vocabulary selection effectively distributed by proficiency levels. This outcome would result from an efficient process of vocabulary input selection distributed by semantic fields and articulated into instructional modules (levels) suited for time- and retention-effective pedagogic strategies.

The description of levels of vocabulary input in progression looks very much like the description of a curriculum using a structured and well-distributed inventory of potential content. Moreover, syllabus design implies incorporation of a segmented vocabulary inventory into instructional units, the sequence in which the units are to be learned, and thus a narrower definition of content. Such inventory, resembling those found in the French *Le Niveau Seuil* or the Council of Europe vocabulary profiles, would contain vocabulary material needed to implement the "Can-Do" approach outlined in the CEFR and would be used by course designers and evaluators alike in preparing learning or test materials.

The importance of accurate lexical selection and distribution has to do with efficiency. Language learners should be able to have learned a critical mass of vocabulary for each intended level, and certainly by the time they start advanced studies or intend to work professionally. Moreover, French internet corpora might be as representative as the CEFR vocabulary or more of current French writing and speech, and allow a more definite answer to the question of stable core vocabulary needed by beginning and intermediate learners of French as a foreign language.

Three main approaches have been used to select and distribute lexical profiles for language teaching: (a) strict frequency studies based on general or on specific professional sources, such as in the Frequency Dictionary of French (FDF) (Lonsdale & LeBras, 2009), (b) needs analysis in general or in behalf of specific target groups such as the ones conducted by Council of Europe experts, and (c) a combination of the two preceding approaches, such as in *Le*

Français Fondamental 1 and *Le Français Fondamental 2* (Gougenheim et al., 1964). In theory all three approaches should lead to similar or at least comparable results since they all aimed at defining lexical content needed to reach the advanced proficiency threshold.

This study investigated to what extent these approaches correspond, how to explain discrepancies, and which recommendations could be given to optimize the selection of vocabulary per level. The research compared methods of selection, profiles composition, and distribution by progression level to identify problems and challenges, and to suggest remedial approaches. Previous research that showed the importance of lexical content in preparing learners of foreign languages for advanced studies, and that examined what is already known about the purpose and use of European vocabulary profiles were reviewed.

Purpose Statement and Research Question

The purpose of this study was to analyze and substantiate the content of the CEFR French vocabulary profiles by answering the question: To what extent do the French types contained in the CEFR French vocabulary profiles overlap with the most frequent types of the FDF and the French Gigaword Corpus (FGC)? Figure 1 illustrates the scope of what the research question intended to identify. Note that the diagram used here and later in the text to depict the degree of overlap is not to scale.

In order to perform this comparison, the four overlapping sections were quantified and described: (a) core, (b) intersection of FDF and CEFR, (c) intersection of FGC and CEFR, and (d) intersection of FDF and FGC. In addition, the three non-overlapping sections were: (a) data unique to the CEFR, (b) data unique to the FDF, and (c) data unique to the FGC. They were evaluated overall, by proficiency level, and by frequency ranking. Frequency rank-order correlations were performed on the FDF and FGC corpus-based resources. The results of this

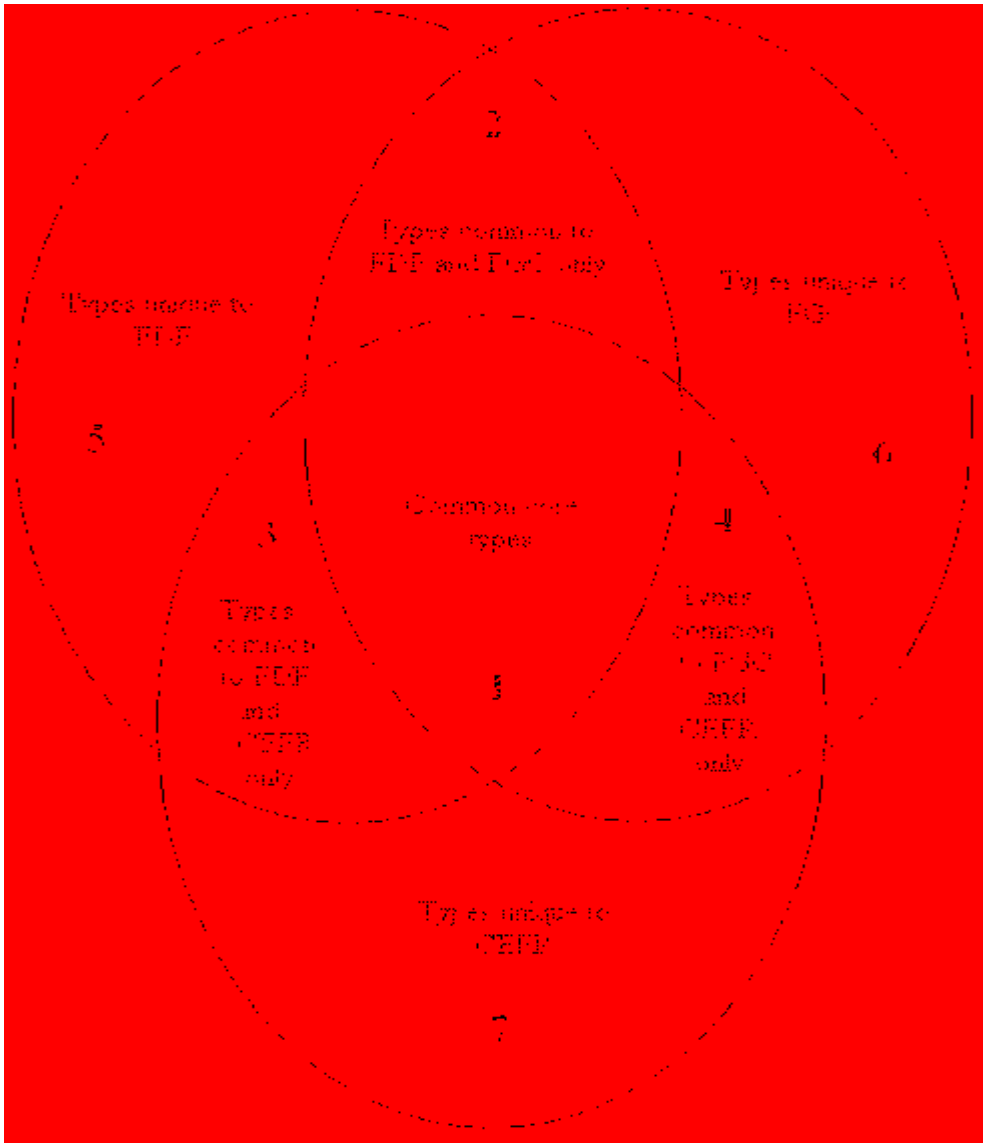

Figure 1. Overlap of the CEFR, FDF, and FGC resources

analysis provide the basis for recommendations made about vocabulary selection and organization for general French coverage at lower proficiency levels and French vocabulary sequencing for beginning and intermediate learners.

In summary, the purpose of this study was to substantiate the content of the CEFR French profiles since they are considered a reference for French teaching, learning, and evaluation. The new knowledge contributed by this study is adding insight into the lexical content of the CEFR French profiles, what they cover, and what they do not cover in relation to most frequent types of the French language that could represent between 80 to 95% of French written or oral texts. The study's emphasis on frequency is not precluding the teaching and learning of other task- or context-dependent words.

Beneficiaries

There are many stakeholders in this issue, not only in Europe but also in other places where French is being learned and taught. These include learners, teachers, instructional designers, publishers, evaluators, linguistic researchers, administrators, and taxpayers. They will all benefit from a usage-based tested body of contemporary vocabulary.

Major Stakeholders. The major beneficiaries of this study are learners and teachers. Learners invest resources of time, money, and effort and expect to be able to use their linguistic knowledge successfully in academic and professional environments. They do not have resources to waste when learning a language they need for advanced French academic purposes or to work in a French professional/occupational environment. Determining what specific French vocabulary should be known for the purpose of being functional at an advanced level will create focus and facilitate, organize, simplify, and accelerate the learning process. It will allow the transfer of this process to other languages and will fill in gaps that would otherwise subsequently

handicap learners and teachers. Teachers who are unlikely to have the time or training to conduct evaluation studies, will still need a reliable source of information and a base against which they can measure student progress. This evaluation will increase the quality of French language instruction by pointing out lexical areas not covered by existing vocabulary profiles, thus benefitting language teachers.

Other Stakeholders. In addition to teachers and learners, there are other stakeholders who would benefit from this knowledge. Instructional designers need to have a sensible and explicit rationale for the progressive sequence they choose in presenting their linguistic information. Publishers aim to endorse state-of-the-art teaching methodologies enhanced by current and adequate content. Evaluators spend considerable amounts of time developing test instruments that determine learning outcomes, and need to be confident that they are working with a reliable and valid vocabulary base. Linguistic researchers often lack the funding to do extensive research themselves, but will appreciate the feedback provided by their colleagues' evaluation research. Administrators at various levels, fund such research projects, and hope they will make a difference and help individuals achieve educational, professional, social, and economic goals they could not otherwise reach. Taxpayers will have added incentives to contribute to education budgets if they have proof of the efficacy of the methods used to accomplish language learning outcomes. All these interested parties will gain from a tested French vocabulary base which is more representative of current usage.

Definitions

This section is defining and clarifying key concepts used to explain the problem at hand and the proposed solution. There are many acronyms used in this study.

- *CEFR* stands for *Common European Framework of Reference* (for languages).

- *FDF* stands for *Frequency Dictionary of French*.
- *FGC* stands for *French Gigaword Corpus*.
- *FFL* stands for *French as a Foreign Language.*
- *EFL* stands for *English as a Foreign Language.*
- *L1, L2* stand for a person's *first (native)* and *second language* respectively.

Lexical competence is an expression used to describe an aspect of foreign language competence. It has been defined as the knowledge of, and ability to use, the vocabulary of a language (Council of Europe, 2001). This vocabulary plus all its elements is also called the lexicon.

The communicative approach, also known as communicative language teaching (CLT), is a general expression which emphasizes the conveying of meaning and interaction as the process and the goal of language learning.

CEFR *level(s)* refer to degrees of language proficiency as described in the Common European Framework of Reference for languages (in French, *niveau(x)* du Cadre Européen Commun de Référence pour les langues (CECR)) evaluate the autonomy and independence a learner exhibits in the use of a language. The CEFR divides learners into three broad categories which are subdivided into six levels (see Table 1).

Lexical progression is a level-related concept. This concept refers to language development stages and movement towards language mastery.

Lexical coverage is a term used in foreign language pedagogy to describe the percentage of words the reader of a certain text understands immediately.

The terms c*orpus* (singular) and *corpora* (plural) are used herein to mean an analyzable group or collection of electronic texts, loosely or tightly structured as an identifiable whole for a

given purpose, such as the study of some aspects of language and linguistic analysis, and from which vocabulary lists can be derived.

Table 1

CEFR Descriptive Scales for Language Proficiency Levels

Category	Level	Description	Other French Labels	Other British English Labels
A: Basic Learner	A1	Beginner	Niveau A1	Breakthrough
	A2	Elementary	Niveau A2	Waystage
B: Independent Learner	B1	Pre-intermediate	Niveau B1	Threshold
	B2	Intermediate	Niveau B2	Vantage
C: Proficient Learner	C1	Upper intermediate	Niveau C1	Effective Operational Proficiency
	C2	Advanced	Niveau C2	Mastery

The term *lemma* as used herein means the base form of a word representing all inflections listed by part of speech. A detailed example is found in Table 3 of Chapter 2. A lemma, similar to lexical groupings, includes a word baseform and all its inflections, but not its common transparent derivations.

The term t*ype* as used herein means the individual spelling of a unique string of characters separated by whitespace or punctuation, and neutralized for capitalization. For instance, *donne-moi* or *l'homme* count as two types.

French vocabulary is limited to the presence of French *types*, one-word units, such as nouns, verbs, adjectives, and adverbs, and function words as well, also called grammatical elements, belonging to closed word classes such as articles (e.g., *le, un*), quantifiers (*quelque, tout, plusieurs*, etc.), demonstratives (e.g., *ce, cet, ces*), personal pronouns (e.g., *je, tu, il*), possessives (e.g., *mon, ton, sa*), prepositions (e.g., *dans, à, par*), adpositions, conjunctions (e.g.,

et, mais, si), auxiliary verbs (e.g., *être, avoir, aller*), question words and relatives (e.g., *qui, que, où*), interjections (e.g., *ouf*, *mince*), or particles (e.g., *ne*). In this study no multi-word units were considered.

Delimitations

Given the complexity involved in analyzing morpho-semantic units, the chosen unit of analysis for this research was the *type*. The development of morphological knowledge, i.e. the learning of *types*, is a lifetime pursuit. Beginning L2 learners have difficulty establishing morphological connections. The numerous inflections of the French language take time to learn. This learning should preferably happen early on in foreign language acquisition.In addition, given current technology, one of the primary source of this study, the FGC resource, with close to a billion words, would have been extremely difficult to lemmatize. *Types* were chosen as an alternative to *lemmas* that would allow comparable categorizing, counting, ranking, and analyzing. Moreover, t*ypes* added the granularity needed to see which French word forms were more prolific than others. However, another primary source of this study, the FDF resource, only came with frequency and rank information for lemmatized data, thus no frequency ranking for FDF inflections. Given the previous definition of the concepts of *type* and *lemma*, the analysis conducted here focused on form. A meaning-based analysis was excluded due to the fact that the technology needed for such is still in its infancy (Gardner, 2007), and adequate French semantic taggers are not available at this time.Beyond the scope of this study lies the comprehensive question of "what vocabulary should be included in a content-based advanced French as a foreign language profile?" This study will only begin to address the issue by focusing on what general lexical threshold can be reached with current instructional design.

Chapter 2: Review of Literature

This study compared three primary vocabulary resources with distinct features. The CEFR French vocabulary profiles were produced by language acquisition experts on the basis of communicative tasks. On the other hand, the FDF and FGC were produced by linguists using very large to mega electronic corpora. A characterization of these sources is presented in order to better determine their origin, what they include, and how they have been used so far. In addition, the review discusses what number of words, at what rate, and which words are needed to function at a given level; what is the best method to count lexical units for instruction; and what is the best use of corpora to answer the research question.

Key Concepts in Regard to Vocabulary Teaching and Learning

Certain terms such as lexical competence, the communicative approach, CEFR proficiency levels, lexical progression, lexical coverage, corpus, lemma, and type are consistently used in linguistics and language acquisition terminology. These key concepts were briefly defined in Chapter 1 and they are further explained here.

Lexical competence. Lexical competence has been used to describe foreign language competence. It is defined as the knowledge of, and ability to use, the vocabulary of a language (Council of Europe, 2001). Even though a lexicon has real complexities, lexical competence has been described with few measurable dimensions attached to the lexicon in its entirety as opposed to individual lexical items (Meara, 1996). These dimensions were identified as lexical size, i.e., size of vocabularies with the rate at which they grow and factors affecting this growth. It also included lexical organization, including the ability to produce native-like associations with foreign language words (Deese, 1965; Kiss, 1968; Richards, 1976). Until the late 1990s, lexical competence did not have a definition distinct from grammatical competence and was considered

part of it, since the main emphasis was then placed on singling out communicative competence (Canale & Swain, 1980; Meara, 1996). There was, until then, a lack of information about the role vocabulary plays in language (Zechmeister, D'Anna, Hall, Paus, & Smith, 1993).

The communicative approach. The phrase "communicative approach" was coined in the 1970s to describe an innovative trend focusing on notions and tasks in foreign language teaching. Historically, it followed the behaviorism-based audio-lingual approach used during and after World War II, and paved the way for the notional syllabus (also called functional or notional-functional) where notions and functions, instead of grammatical structures, became the way to organize the language curriculum. A notion is equivalent to a context for communication, e.g., *working*, *playing*, *shopping*, *traveling*. A function is equivalent to a purpose for interacting in a specific context, e.g., the notion or context of *working* would necessitate several language functions such as greetings, asking about tasks, bargaining, writing a response to a client. The communicative approach is encouraging initiative, cooperation, and role-plays between learners as well as grammar and pronunciation activities. However, the approach is focusing more on task outcome, language fluency, and student confidence than on accuracy of language forms.

CEFR proficiency levels. CEFR users including the French vocabulary profiles authors, were to specify which lexical elements the learner would need to know and use at a given level, and how these elements were selected and ordered (See Table 2). This description of levels of vocabulary input, as stated earlier, are looking very much like the description of a sequenced curriculum building on internal interrelations and continuities among major units of instruction intended to improve learning (Decoo, 2011). This kind of description is informing a planning syllabus, providing a practical prospective structured inventory of potential content distributed into instructional units.

Table 2

CEFR Descriptive Scales for Vocabulary Knowledge Range and Control

CEFR level	Vocabulary Range Descriptors	Vocabulary Control Descriptors
A1	Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.	No descriptor available
A2	Has a sufficient vocabulary for the expression of basic communicative needs; has a sufficient vocabulary for coping with simple survival needs.	Can control a narrow repertoire dealing with concrete everyday needs.
B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.	Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.
B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.	Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication
C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.	Occasional minor slips, but no significant vocabulary errors.
C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.	Consistently correct and appropriate use of vocabulary.

The specifications of aims, selection, and grading within the selections are defining content more precisely, and determining what is needed for narrowed progressive lists (i.e., the planning syllabus) that could be used for the construction of a course. Task-based vocabulary inventories, such as those found in the French vocabulary profiles, are containing source material for prospective methods to implement the "Can-Do" approach outlined in the CEFR. They are to be used for the preparation of learning materials and incorporated into syllabus design, since syllabus design allows decisions regarding segmentation into instructional units, and the sequence in which they are to be learned.

Lexical progression. Lexical progression refers to language development stages and movement towards language mastery. Some aspects of vocabulary proficiency are analyzed longitudinally at different stages of the learning process, for example between the last grade of high school and the first term of university studies within a language. For instance, the American Council on Teaching of Foreign Languages (ACTFL) is suggesting that

> college freshmen should have attained, by the end of high school, the ability to listen, converse, read, and write in the target language with sufficient basic skill, vocabulary, accuracy, and cultural awareness to communicate needs in everyday situations in a culturally appropriate way. (California State Department of Education, 1986, p. vii)

The concept of progression is presupposing prior knowledge of the language and pre-assessment before instruction, pre-assessment at every junction of instruction, i.e., between primary and secondary, secondary and higher, higher and continuing education and every articulation in between in order to achieve continuity throughout the process of language learning, and language instruction coherence between levels.

Lexical coverage. Lexical coverage is representing the number of words known by the learner in the text, multiplied by 100 and then divided by the total number of words in the text (Nation, 2001). Below 80%, reading comprehension is almost impossible (Hu & Nation, 2000). Ninety-five percent coverage is the point at which learners can read without the help of dictionaries (Laufer, 1989). Research (Hu & Nation, 2000) showed that 98% should be the desired coverage for reading comprehension, equating to a maximum of one unknown word in 50. The concept of coverage came into play to determine how many words are needed for fluent reading and understanding of a language, and of a text in particular. Coverage was identified through frequency studies. Frequency studies have been the most rapid way to identify what to teach in order to meet learners' general linguistic needs.

Corpus. A corpus is defined as an analyzable collection of electronic texts. Corpora (plural) can be researched to answer questions about the "prosody, lexis, grammar, discourse patterns or pragmatics (of a language)" (Kennedy, 1998, pp. 3-4). As an example, a corpus according to the present research could be a compilation of electronically recorded French newswires of any length. The compilation could as well include books of any genre such as French literature, technical, legal textbooks and others, published between the 17th and 21st century with the purpose of giving quick access to French books. Examples of corpora include structured electronic texts used to produce frequency dictionaries or dictionaries of oral or written language, and unstructured online corpora, such as Google N-grams, used to produce vocabulary lists and word frequencies. The building blocks of a corpus are strings of characters separated by spaces generally described as morpho-semantic units such as single or multiword units. These units are typically counted (e.g., by token, type, lemma, word family, and the like),

ranked (e.g., by frequency, notions, themes), analyzed (e.g., morphologically, semantically, and the like), and categorized (e.g., by level, difficulty, and the like).

Lemma. It is typically a common content word, for instance, in the form of a noun, a verb, or an adjective. The example of the verbal lemma *work* (*travailler* in French) and all the word forms attached to it follows (see Table 3). A lemma might have more than one meaning. For example, the English baseform *work* has over 40 meanings, but only consists of a single word unit. In contrast, a lexeme has only one single meaning regardless of the number of words it contains, and thus often consists of multi-word units, such as idiomatic expressions, e.g. *to kick the bucket* meaning to physically die.

Table 3

English and French Examples of a Lemma

Lemma	Word baseform and inflected forms represented
work (English verb)	*work, works*, *worked*, *working*, *wrought*
travailler (equivalent French verb)	*travailler, travaille, travailles, travaillons, travaillez, travaillent, travaillais, travaillait, travaillions, travailliez, travaillaient, travaillai, travaillas, travailla, travaillâmes, travaillâtes, travaillèrent, travaillasse, travaillasses, travaillât, travaillassions, travaillassiez, travaillassent, travaillerai, travailleras, travaillera, travaillerons, travaillerez, travailleront, travaillerais, travaillerait, travaillerions, travailleriez, travailleront, travaillé, travaillée, travaillés, travaillées, travaillant, travaillante, travaillants, travaillantes*

It is noteworthy to observe that the French/English lemma ratio in this example was 42:5. Moreover, disambiguation of part of speech becomes necessary when the same lexical form could, for instance, count as (a) a noun or a conjugated verb, e.g., *juge* which in French could be a *judge* or *judges*, third person singular of the indicative or the subjunctive present tense of the verb to judge; (b) a noun or an adjective, e.g., *américain*, which in French could be *an American*,

the noun or the adjective *American*; or (c) an adjective or a participle, e.g., *loué* (meaning *rented* or *praised*) with the possibility of the word *loué* being used as in the English a *rented* car or as in "The apartment is *rented*."

Type. A type is the individuela spelling of a unique string of characters separated by white space or punctuation. Examples of types related to the lemma *work* (*travail* or *travailler* in French) are presented here in Table 4. As with a *lemma*, a *type* might have more than one meaning, but is only consisting of a single word unit.

Table 4

Types for the Words Work and Travailler

Count	Types related to the English word: *work*
10	*work, works*, *worked*, *working*, *workings, wrought, worker, workers, workable, unworkable*
	Types related to the French word: *travailler*
53	*travailler, travaille, travailles, travaillons, travaillez, travaillent, travaillais, travaillait, travaillions, travailliez, travaillaient, travaillai, travaillas, travailla, travaillâmes, travaillâtes, travaillèrent, travaillasse, travaillasses, travaillât, travaillassions, travaillassiez, travaillassent, travaillerai, travailleras, travaillera, travaillerons, travaillerez, travailleront, travaillerais, travaillerait, travaillerions, travailleriez, travailleront, travaillé, travaillée, travaillés, travaillées, travaillant, travaillante, travaillants, travaillantes, travailloter, travailleur, travailleuse, travailleurs, travailleuses, travail, travaux, travaillisme, travaillismes, travailliste, travaillistes*

French Lexical Resources

Three lexical resources are used in this study. The first one is referred to as the CEFR French vocabulary profiles since they were inspired by the Common European Framework of Reference for languages document. Then, the second and third resources emanate from the Frequency Dictionary of French (FDF) which is based on a 23 million corpus and the French Gigaword Corpus (FGC) which contains close to a billion word of French newswire text.

CEFR French vocabulary profiles. The makeup of the CEFR French vocabulary profiles can be best understood with two quotes from the panel of expert authors of the French CEFR vocabulary profiles (from their most recent B1 level profile 2011 publication) that explain their methodology and showed that their vocabulary selections have not been tested against large corpora.

> . . . as in *Levels for French* already published (B2, A1.1, A1 and A2), this document of reference for teaching proposes to identify forms of French likely to correspond to descriptors which characterize the level of reference equivalent to the *Framework*. It originates from several sources which tend to legitimize these choices: the expertise of the authors, the collective expertise of decision-makers regarding teaching and evaluation programs, research findings in French acquisition, the knowledge of discursive genres. Any specific local choice might be arguable but we did make sure that the whole was coherent with the CEFR descriptors... (translated from Beacco et al., 2011, p. 6)

Authors of the French vocabulary profiles further explained their selection criteria and the nature of their vocabulary inventories by stating

> . . . to base this specification of B1 on criteria which would be objectivized and as germane as possible, we took into account:
>
> - essentially the *Framework* descriptors;
> - the knowledge considered established on learners' interlanguages and on acquisition sequences of French;
> - common teaching objectives, in particular the morphosyntactic materials offered for learning in beginners' manuals;
> - collective experience of teachers and evaluators;

- examples of learners production known to be associated to one level or another, and related to *Framework* criteria only (in particular, such samples produced for the French language). (translated from Beacco et al., 2011, pp. 14-15)

French vocabulary profile authors recognized the value of corpus linguistics methodology when acknowledging its use by their English colleagues for the CEFR English profile based on learner corpora and supplemented by megacorpora, e.g. the Cambridge English Corpus. The authors stated their lack of adequate financial resources prevented them from using this methodology (Beacco et al., 2011). Instead they relied on established pedagogical knowledge, which concurs with Hulstijn's assessment of the CEFR, and guided their selection methodology when the later states that "the CEFR empirical base consists of judgments of language teachers and other experts with respect to the scaling of descriptors" (Hulstijn, 2007, p. 665). This reality becomes one of the main reasons for conducting this study in order to test CEFR French vocabulary content against usage-based corpora evidence.

Development of the CEFR. The CEFR French vocabulary profiles were produced over a 6-year span. The highest language proficiency level B2 (intermediate) was published first in 2004, and the lowest, level A1 (beginner) three years later. Level A2 (elementary), was the next level to be published a year later, and finally level B1 (pre-intermediate) in 2011. Their authors were French native-speakers addressing the needs of non-native French learners. Lexical units contained in the French profiles are heterogeneous, i.e., lemmas, types, or multi-word units. Quantitative lexical cumulative input ranges from close to 1,000 lexical units of instruction at the lowest proficiency level to about 6,500 lexical units for instruction at the intermediate level. The vocabulary selections were structured after the notional task-based CEFR. An appreciation of the French profiles' features was gained by understanding that they came about through the means of

European language policy spanning over 40 years, and the prevailing communicative approach for language teaching and learning. The French vocabulary selections were hereafter put in historical and theoretical context.

The Council of Europe language policy panel work started in the early 1970s with *The Threshold Level* (Van Ek, 1975) and *Un Niveau-Seuil* (Coste, Courtillon, Ferenczi, Martins-Baltar, & Papo, 1976), and eventually led in the early 2000s to the adoption of a common proficiency scaling system, i.e., the CEFR, usable for any European language. Initially, European experts and researchers were assigned to assess the needs and personal objectives of language learners so specific corpora could be put together for each language, and vocabulary content specified to reach the "lowest level of general foreign language ability to be recognized in a unit-credit system" (Van Ek,1975, p. 7). This model became the basis for other European language systems, such as German, Spanish, and French. *The Threshold Level* is a specification for minimal general communication proficiency in a foreign language, and it implied proficiency levels above and below the specified threshold.

Thanks to French financing of Council of Europe language policy, and, for the first time in language education, multidisciplinary research teams were formed to analyze the needs of a diversified pool of language learners. The threshold content specification was to serve as a general reference for learners seeking to obtain "minimal" linguistic competence, and was designed with a good amount of pedagogical flexibility and lexical content variability. It is focusing on the varied linguistic needs and objectives of five important FFL learner groups in Europe: (a) tourists and travelers with basic linguistic needs, e.g., to eat, find a hotel, or ask simple questions; (b) migrant workers and their families needing to successfully integrate themselves in society and the workplace; (c) experts and professionals needing FFL but staying

in their home countries with more varied needs, e.g., read specialized literature, talk to foreign colleagues and write them letters; (d) high schoolers and university students with linguistic needs similar to tourists and professionals, e.g. understand scientific literature, meet job description requirements, or be able to communicate when traveling; and (e) teenagers attending school needing to find motivation to start and keep learning a foreign language. This work led to a new definition of language learning based on behaviors appropriate to situations in which learners might find themselves and on anticipated speech acts learners would have to perform in everyday situations.

Almost three decades later, other Council of Europe initiatives went further in the development of proficiency levels and the specification of vocabulary content. The program "Language Learning for European Citizenship" between 1989 and 1996 led to intergovernmental work on the "Transparency and Coherence in Language Learning in Europe: Objectives, Evaluation, Certification" in the early 1990s with the objectives to improve the recognition of language qualifications and help teachers co-operate.

This effort allowed the further development of levels of language proficiency and the creation of a "European Language Portfolio" – a certification in language ability usable across Europe; it also led to the European Union Council Resolution in 2001 to recommend the use of the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR, Council of Europe, 2001) to set up systems of validation of language ability. The driving force behind the decision to produce the CEFR was according to Trim

> ...the need for the portability of the [language] qualifications on offer, whether for the benefit of individual learners and providers, or those, such as ministries, employers and

> university authorities, called on to interpret qualifications from diverse sources and to make administrative decisions on that basis. (2011, p. 10)

The synergy that flowed from European collaboration led to the production of related language profiles.

European vocabulary profiles. Since 2001, after the publication of the CEFR, massive national efforts in nearly all European countries have led to the development of vocabulary profiles (referred to as *Référentiels* in French). These profiles are recommended minimal lexical lists of task-based general and specific vocabulary notions. The entries of these lists were selected to fit proficiency levels, as described in the CEFR (See Definitions in Chapter 1). They have as one of their main objectives to measure how well students know their foreign languages. They were also intended to support the process of presenting lexical learning and testing materials which should prepare learners to reach advanced lexical competence. The CEFR is specifying what vocabulary might be attached to each level as defined in Table 1. The CEFR is also describing what a language learner should be able to do at each proficiency level in reading, listening, speaking and writing. The descriptors are applying to all European languages and, thus, define very generally what a learner should be able to accomplish at each level. Terms such as *familiar everyday*, *very basic*, *frequently used*, and *wide range* qualify the lexical content that is to be learned at various levels. These terms are hinting at vocabulary quantity and quality, but remain unquantified and undefined as Table 5 illustrates. Authors of the booklet *Using the CEFR: Principles of Good Practice* noted that the Framework is a "central point of reference open to amendment and further development" (ESOL Examinations, 2011, p. 2) for teaching, learning, and assessment; it is "not language or context specific" (p. 6). It did not attempt to list vocabulary but in order to use it in a meaningful way, developers must elaborate its contents.

Table 5

Proficiency Level Descriptors

Level	Description
A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.
C2	Can understand with ease virtually everything heard or read. Can summaries information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.

This might have included establishing which vocabulary occurs at a particular proficiency level in a given language. The CEFR guidelines are suggesting, for instance, that an A1-level user “Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.” (2001, p. 112) The A2-level learner “Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.” (p. 112)

Under the CEFR section 6.4.7.2 “*Size*, *range*, and *control* of vocabulary”, users can determine "what size of vocabulary (i.e. the number of words and fixed expressions) the learner will need to control" (p. 150). Under "Lexical selection," the CEFR (2001) is clarifying that authors of materials have a number of options:

> [1] to select key words and phrases a) in thematic areas required for the achievement of communicative tasks relevant to learner needs, b) which embody cultural difference and/or significant values and beliefs shared by the social group(s) whose language is being learnt;
>
> [2] to follow lexico-statistical principles selecting the highest frequency words in large general word-counts or those undertaken for restricted thematic areas;
>
> [3] to select (authentic) spoken and written texts and learn/teach whatever words they contain;
>
> [4] not to pre-plan vocabulary development, but to allow it to develop organically in response to learner demand when engaged in communicative tasks. (pp. 150-151)

The CEFR guidelines are no doubt being very flexible. Authors of the existing French CEFR profiles have clearly opted for the first choice in the paragraph cited above. They have used it as their major selection procedure. This approach leans on notions and functions. The CEFR chapter on assessment is confirming this preference (CEFR, 2001).

The CEFR is stating that content specifications produced for over 20 European languages can be seen as ancillary to the main Framework document. They are offering examples of a further layer of detail to inform test construction for Levels A1, A2, B1 and B2. They are seeming to suggest that "content specifications" in the "ancillary" publications are strongly recommended "examples" of what instruction and test designers should include in their syllabi and tests (p. 179). These CEFR statements are not helping decide whether word selection for a given level should be made within a CEFR-approved inventory, or what level of coverage of an inventory will match with the level, even if users are free to develop their own separate inventory to meet content specifications.

The prior observations made regarding the CEFR are an additional reason for investigating how CEFR French profiles compare with frequency data of large to mega-corpora, and they are showing that CEFR authors did not have a core vocabulary foundation in mind. Moreover, the fact that the CEFR is not being content specific for any of the European languages makes this study relevant to help determine the lexicon related to a level.

Hybrid vocabulary selection approach. Council of Europe language policies of the 1970s coincided with the communicative approach trend for foreign language teaching and learning. This trend was actually born as a reaction to then-prevailing audio-lingual and audio-visual didactic methods, and a response to the linguistic needs of the European community. The communicative approach and a hybrid version of the notional syllabus based on "Can-Do" tasks have inspired the work of European expert panelists who promptly adopted this new way of designing instructional input which continues to this day. The notional syllabus has indeed become the preferred alternative to the formal / structural / grammatical syllabus and can be seen in foreign language instructional design work of Council of Europe commissioned experts.

Seminal works found in the *Notional Syllabuses* by Wilkins (1976), in *The Communicative Approach to Language Teaching* edited by Brumfit and Johnson (1979), in *Explorations in Applied Linguistics* by Widdowson (1979), in the article *Options for vocabulary learning through communication tasks* by Newton (2001), task-based instruction by Skehan (2003) have laid the foundation for thinking about foreign language instructional design in the past 40 years.

Notions and tasks are two classifying concepts which have allowed subjective vocabulary selection on the part of teachers and learners alike. They influenced even more than actual definitions of curricular syllabi the learning content specification, the selection, and the sequencing of foreign language content (Allwright, 1984; Prabhu, 1987; Stern, 1983; Yule, Powers, & Macdonald, 1992). Notions and functions are linguistic terms that describe communicative activities. A notion, in the linguistic context, is referring to the situation in which the communication takes place and to the ideas and the information that need to be sent and received. Functions are specifying the kinds of interaction and aims the language user wants to accomplish through and with the language. "To exchange information" is an example of a notion where communication will take place and information will be exchanged. The functions derived from this notion could be "to give information" (ex.: *Paul a l'air malade*/*Paul seems sick*), "to express surprise" (ex.: *C'est surprenant*/*It's surprising*.), or "to express ignorance" (ex.: *Je ne le savais pas*/*I didn't know it*.) General notions are referring to words that belong to several referential fields, or possibly all, such as words related to the notions of time, space, and quality. Specific notions, in contrast, are referring to words that represent only one specific referential field.

The general and specific notions framework, initially used by Trim and his English colleagues, then adopted by their French colleagues, Coste et al. (1976) and, more recently,

Beacco et al. (2004, 2007, 2008, 2011) is shown in Table 6 with French vocabulary selections (and their English translation) exemplifying the notional/task-based framework descriptions and categories.

Functional syllabi properties identified by Breen (1987) help clarify decision-making criteria used for notional task-based vocabulary selection. Vocabulary knowledge is being prioritized and centered on speech acts in the context of social activities or events in order to negotiate, interpret, and express meaning. Language, as a means for accomplishing tasks, is being given priority over linguistic knowledge in itself with the wish to enable learners to use language - practically from the beginning of their learning - in order to achieve interpersonal and social goals. The functional syllabus categorizes main types of language purposes in sets and subsets. It specifies how the functions may be accomplished through various language options, and from general, more common sets of functions to more specific and varied functions.

The sequencing of what is to be learned from tasks is cyclic (in relation to how learners move through tasks), and problem-generated (in relation to the ongoing difficulties the learners themselves discover). As the learner is progressing, and tasks are requiring more and more linguistic competence, there is a sequence of diagnosis and refinement. Since learners have to identify learning problems or difficulties, they have to prioritize problems and the order they may be dealt with, and identify the appropriate learning tasks which will address the problem areas.

CEFR profile authors are using a hybrid approach to selecting lexical units. The first underlying concept of this approach is centering on general and specific semantic notions. The second concept is language tasks.

Table 6

Communicative Task-Based Framework of General and Specific Notions

I. General Notions	Examples (French)	Examples (English)
I.1 Existence	*présence, devenir, optionnel, ne*	*presence, become, optional, not*
I.2 Time	*retard, soudain, plus tard*	*delay, sudden, later*
I.3 Space	*lieu, local, où*	*location, local, where*
I.4 Quantity	*cent, partie, quatrième*	*hundred, part, fourth*
I.5 Quality	*carré, pointu, mouillé*	*square, pointed, wet*
I.6 Relations	*comparaison, similaire, le mien*	*comparison, similar, mine*
II. Specific Notions	**Examples (French)**	**Examples (English)**
II.1 The person	*nom, immortel, attentivement*	*name, immortal, attentively*
II.2 The house	*palais, vivre, non-meublé*	*palace, reside, unfurnished*
II.3 Around the house; nature; weather; seasons and celebrations	*terrain, se jeter, terrestre*	*ground, flow into, earthly*
II.4 To go somewhere	*départ, embouteillage, longer*	*departure, traffic jam, go along*
II.5 To eat and drink	*recette, mûr, préparer*	*recipe, ripe, prepare*
II.6 Trade and errands	*acheteur, pièce, vêtu*	*buyer, coin, clothed*
II.7 Public and private services	*timbre, raccrocher, au secours*	*stamp, hang up, help*
II.8 Hygiene and health	*corps, hôpital, féminin, fatiguer*	*body, hospital, feminine, tire*
II.9 Physical notions	*lentille, regarder, en avant*	*lens, watch, forward*
II.10 Work	*mi-temps, faire, mèl*	*part-time, do, email*
II.11 Hobbies	*loisir, hifi, vacances*	*leisure, hi-fi, holidays*
II.12 Human relations	*voisin, questionner, message*	*neighbor, question, message*
II.13 Current events and daily activities	*cas, juger, arrêter*	*case, judge, arrest*
II.14 Education	*lycée, classer, former*	*high school, rank, train*
II.15 Language	*parlé, parole, bavard*	*spoken, word, talkative*

Limitations of the task-based notional syllabus. The task-based approach aimed at relating content to how content may be worked upon in order to have a catalytic effect on language acquisition, and match native language acquisition processes. It also rested on the principle that communicating about communication is itself a great trigger for language learning. The evolution of the concept of *task* over the 1980s and 1990s has led to strong and weak variations of task-based syllabi, where the term *task* has finally been defined as "anything the learners are given to do (or choose to do) in the language classroom to further the process of language learning." (Williams & Burden, 1997, p.167) According to Nunan, selecting, sequencing, and integrating tasks is the crux of notional task-based syllabi.

> The essential problem to be solved, ... is how to achieve a rational articulation in selecting, sequencing and integrating tasks so that the curriculum is more than an untidy 'rag-bag' of tasks which, while theoretically motivated in psycholinguistic terms, are unrelated to each other and disconnected from the learner. (1993, p.56)

Some of the research on methods for selecting tasks also pointed out their connections with vocabulary selection and the importance of classifying tasks according to their difficulty so that task selection and grading can be more effective (Skehan, 1998). Researchers have identified factors influencing task selection: (a) task complexity (number of steps involved, complexity of instructions, cognitive demands, quantity of information) which influence language demands; (b) language input recognition (making sense of how the language is organized and structured); (c) sequence of input; (d) explicitness of input; (e) type of input; (f) precision of input; (g) the amount and type of information provided; (h) the degree of abstractness of the concept dealt with in the task; (i) how much information is contained in the input; (j) the vocabulary used; (k) the genre, discourse, structure of items in a text; and (l) prior information (Anderson & Lynch, 1987;

Brindley, 1987; Candlin, 1987; Candlin & Nunan, 1987; Nunan, 1989; Prabhu, 1987; Robinson, Ting, Urwin, 1996; Skehan 1992). Criteria for task selection researchers have identified do influence vocabulary selection, and, in turn, vocabulary learning.

Candlin's guideline proposal for task selection showed that language learning under the communicative approach has not been conducted in a well-ordered fashion (1987). He suggested, for instance, that (a) one-way tasks should precede two-way tasks; (b) static tasks should precede dynamic tasks; (c) tasks in the present time should precede ones using the past or future; (d) easy tasks should precede difficult ones; and that (e) simple tasks (only one step) should precede complex tasks (many steps). Skehan warned that too much focus on meaning during task performance was to the detriment of form and limited vocabulary growth (1996).

The more salient problems related to task-based vocabulary acquisition were that (a) "Natural sequences do not really exist in sufficient detail to be used as the basis for a precise order, nor have they been shown to facilitate learning in a second language situation." (Schinnerer-Erben, 1981, p.11); (b) communication strategies to convey meaning sometimes bypass the learning of word forms (Kellerman, 1991); (c) there are no valid, user-friendly sequencing criteria – one of the oldest unsolved problems in language teaching (Widdowson, 1968); (d) there is no control over number of tasks, types of tasks, and task boundaries; (e) the general tendency is to minimize linguistic forms and the volume of language used by producing only that which is necessary to accomplish the tasks (Seedhouse, 1999); (f) learners interact at the lowest level of explicitness necessary to complete the tasks (Seedhouse, 1999); and (g) task-based syllabi have not been submitted to rigorous, controlled evaluation (Long & Crookes, 1993).

To date, the responsibility to align the vocabulary profiles with learners' needs is resting on vocabulary profile users and learners themselves. Coste has warned that the worst possible use of the French profiles would be to literally take all the words indexed and to consider them as the "scientific" syllabus content to be taught. The only justified use of profile vocabulary content would have to have as a starting point learners' needs and objectives (Roulet, 1977). Again, the French vocabulary profiles are considered by its authors mainly a useful reference and a means of comparison and realignment for language program directors, course designers, and teachers across Europe. They should make use of profile content to produce, evaluate, or analyze didactic materials (new and old) after a careful definition of learners' language needs and goals.

Thus, another reason for this study was that notional task-based syllabi and the vocabulary derived from them end up giving great latitude to learners and teachers for variable subjective content. This content cannot be standardized and it might reflect - without the possibility of a quantifiable evaluation – higher or lower language proficiency.

Intended use of CEFR profiles. With their one decade history, CEFR-defined proficiency levels are now used as yardstick and labels to more clearly identify instructional and evaluation products. They require that European government and educational institutions adjust their language acquisition programs. They are meant to harmonize European language programs, align measurement criteria, and facilitate planning and assessment. The CEFR and instruments derived from the CEFR such as the Europass, the European Portfolio, and Association of Language Testers in Europe (ALTE)-approved assessment tools are being used or made available to students as internationally accredited instruments. The CEFR has been mainly used in the area of language testing. For instance, the ALTE "Can-Do" project developed a simplified

set of 400+ descriptors relating to CEFR levels for language examinations such as the Cambridge EFL exams. Today many more examining boards link their exams to the system.

Weir (2005) discussed problematic limitations of the CEFR for test development or comparability. He stated the CEFR hardly helps identify the breadth and depth of the vocabulary needed to function at the various CEFR language proficiency levels since only general, and sometimes no, guidance is offered in the descriptors with no examples of typical vocabulary associated to them. He recommended that the production of vocabulary profiles for each language be tested and compared so they could supplement the CEFR and give meaning to undefined terms test item writers and item bank compilers have to interpret (Weir, 2005; see also Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu, 2004; Huhta, Luoma, Oscarson, Sajavaara, Takala, & Teasdale, 2002). So the question remained: Was the notional task-based French profile content meant to be used for instruction and learning, or was it not? According to Riley (1982), even though notional syllabus refers to content or teaching material, using its explicit content for teaching and learning was not the intent of CEFR vocabulary profile authors. They do not contain methodological instructions. They were only designed to be works of reference.

> Nonetheless, it has to be recognised that they (the syllabi) have often been used as materials - this is a travesty of the authors' intentions, although the proliferation of "levels" (Threshold, Waystage) can only aggravate the misunderstanding. Attempts to "repair the damage" such as Roulet's "mode d'emploi" for *Un niveau-seuil* are as much a symptom as a cure." (p.98).

Coste et al. (1976) explained that the profiles should not be used as closed and restrictive inventories but as springboards, and thus were minimal lists by design. The credit system, scaled

according to language proficiency, let one assume that the French profiles did offer linguistic content for learning and instruction. The great latitude left to the user for want of methodological instructions encouraged its use as a well-defined content.

If the work of CEFR vocabulary profile users consisted only in evaluating existing or new syllabi, how should they use an instrument that does not delimit its own content? The notional framework failed to be a reference in that in order to evaluate an existing course a user is forced to analyze its content according to its extremely detailed notional categories.

Trim, head of the Council of Europe project in the United Kingdom, confessed from the beginning that the application of the Threshold reference to the needs of specific groups of learners would be a difficult task since much still needed to be learned regarding application methodology (Coste et al., 1976). Without methodological guidance, profile users are thus left to their own devices to make judicious vocabulary choices, and, because of the atomization of language into notions, users are also assumed to know profile content before they use them.

The use of a L2 notional syllabus and the vocabulary attached to it is conditional upon advanced knowledge of the language. This implies that some CEFR profile words might pertain to advanced proficiency levels. Non-native teachers might not have reached those proficiency levels and would not be in a position to teach these words properly.

Breen (1987) identified major questions confronting task-based syllabus designers. One question was particularly relevant to this study. “How might the focusing, selection, subdivision, and sequencing of content become explicit elements within the classroom experience” (p.160)? Even though the *task* as a unit of syllabus design has become an accepted concept, documented research on the communicative approach in the classroom has been insufficient to bring more light into the philosophical, theoretical, psycholinguistic, sociolinguistic and evaluative aspects

of syllabus design, let alone its lexical aspect (Bailey & Nunan, 1996; Canale & Swain, 1980; Legutke & Thomas, 1991; Long & Crookes, 1993; Shaw, 1997). Challenges faced by syllabus designers wanting to develop language proficiency using the CEFR came from several issues. These include the fragmentation by one notion at a time (Crombie, 1985; Widdowson, 1978); no limit to possible notions and functions and overlap very likely (Long & Crookes, 1993); no sound psychological basis to the approach (Cook, 1985); and a basis on reasoning rather than empirical evidence (Brumfit, 1981; Paulston, 1981). Other questions were raised when vocabulary was presented within a notional task-based framework. What vocabulary was the framework leaving out? What methodology should be used to help researchers identify words excluded from these notions and tasks? In what ways did the hybrid syllabus influence vocabulary selection and sequencing by proficiency levels? Why should this vocabulary take precedence over any other vocabulary presented to beginning and intermediate learners? These unanswered questions were yet other reasons for wanting to substantiate the content of the CEFR French vocabulary profiles by means of large corpus-based comparisons.

Frequency Dictionary of French – Core Vocabulary for Learners (FDF). The FDF is the most current frequency dictionary originating from a large and balanced corpus of the French language. The authors of the FDF undertook the project to serve FFL learners and "prepare students of French for the words that they are most likely to encounter in the ‘real world’" (Lonsdale & LeBras, 2009, p.1). It was produced to fill a gap since earlier corpus-based frequency dictionaries of the French language were produced with smaller, more specialized corpora. French dictionaries are numerous as Lonsdale and LeBras point out. Some were based on textual sources of half a million words or less (Henmon, 1924; Juilland, Brodin, Davidovitch, 1970) and some were developed for advanced scholarly purposes (Beauchemin, Margel, &

Théoret, 1992; Brunet, 1981; Imbs, 1971-1994). Some require internet access and subscription such as ARTFL FRANTEXT and TLFI. Some are exclusively in French (Gougenheim, 1958) and others list variable numbers of words without explanation of word selection methods (Lazare, 1992; Buxbaum, 2001).

In contrast, FDF is based on a 23-million-word corpus of French taken from sources of the 1950s or later. The authors stressed its practicality and usefulness to learners of all levels. The top and core 5,000 most frequently used French lemmas, with thematic boxes listing top words by specific topic, were listed in order to access key French vocabulary quickly and easily. The FDF cover page states: "The dictionary provides the users with detailed information for each of the 5,000 entries, including English equivalents, a sample sentence, its English translation, usage statistics, and an indication of register variation." (Lonsdale and LeBras, 2009)

Table 7 details its composition. Its text sampling design, even though it contains French from France and the French-speaking world, is not based on geographical region or demographics; it is based on a balance of genres. Half are oral (11.5 million) and half written (11.5 million). The purpose was to obtain a balanced and more objective representation of the language and, within those registers, to select representative texts of the French language.

Even though frequency dictionary lemmas might not represent every word a learner might need to perform a given task (for example, the word *spoon* (*cuillère* in French) is not present in the frequency dictionary) highly frequent words of the French language are identified and counted. The example of Rolland and Picoche (2008) and their work with the *Trésor de la langue française* (TLF - Treasure of the French Language) corpus can be quoted here. They stated that 907 high frequency French words (repeated over 7,000 times in the corpus) cover 90% of the 170 million word corpus. The crossing of TLF hyperfrequent words (repeated more

than 25,000 times, and predominantly irregular verbs) with the earlier Gougenheim and the later Baudot frequency lists yielded 114 core, generally very polysemous words going from *aimer* to *vrai* (*love* to *true*) which they presented for CEFR level A1 and beyond, but without being able to define under which upper CEFR levels they may be categorized (Baudot, 1992; Brunet, 1987; Gougenheim, 1958; Imbs, 1971).

Table 7

*Composition of the 23 Million-Word FDF French Corpus**

Register	Approximate Number of Words	Type	Sources
Spoken	175,000	conversations	3
	3,750,000	Canadian Hansard	4
	3,020,000	Misc. interviews/Transcript	5
	1,000,000	European Union parliamentary debates	6
	855,000	Telephone conversations	7
	4700,000	Theatre dialogue/monologue	8
	2,230,000	Film subtitles	9
Total	11,500,00		
Written	3,000,000	Newswire stories	10
	2,015,000	Newspaper stories	11
	4,734,000	Literature (fiction, non fiction)	12
	434,000	Popular science magazine articles	13
	1,317,000	Newsletters, tech report, user manuals	14
Total	11,500,000		
Grand Total	23,000,000		

* Table taken from the dictionary (Lonsdale & LeBras, 2009, p.3)

Prior work and applications of corpus and computational linguistics confirm that the lexical content of the FDF 23-million word corpus is a valid lexical resource. It could be used to substantiate and compare CEFR lexical selections.

French Gigaword Corpus (FGC). The French Gigaword corpus was produced in three installments by several authors working for the Linguistic Data Consortium (LDC) hosted at the University of Pennsylvania. LDC is an open consortium of universities, companies and government research laboratories which supports language-related education, research and technology development. It creates, collects and distributes speech and text databases, lexicons, tools, and other resources for research and development purposes. It was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by a grant from the Information and Intelligent Systems division of the United States National Science Foundation.

The FGC represents a growing mega-corpus nearing one billion words of unstructured French newswire text from two major French media agencies, Agence France Press and the Associated Press French Service. These agencies provide news reports used by news organizations such as newspapers, magazines, radio, and television broadcasters. Thus, the FGC resource represents both written and spoken contemporary French. It is presented as electronic text, as seen in the example taken from an LDC product listing in Appendix B.

French Gigaword First, Second, and Third Edition (Graff, 2006; Graff et al., 2011; Mendonça et al., 2009) are part of a sequence of licensed electronic products adding to each earlier version, and constitute a comprehensive archive of newswire text data for information retrieval, language modeling, or natural language processing uses, that has been acquired over about 15 years (between May 1994 and December 2010) by LDC authors for the DARPA GALE Program (Defense Advanced Research Projects Agency Global Autonomous Language Exploitation). Table 8 details the FGC composition.

Table 8

Composition of the French Gigaword Corpus, 3rd Edition

French Gigaword Corpus Source Agencies	Number of tokens (millions)*	Number of documents
Agence France Presse	641.3	2,356,888
Associated Press French Service	221.4	801,075
Total	862.8	3,157,963

* equals the number of whitespace-separated tokens (of all types) after all SGML tags are eliminated

The creation of the French Gigaword corpus involved collection from data sources, i.e. newswire, annotation, management of data, and corpus journal keeping to help replicate or develop similar corpora. It also required converting human-readable into machine-readable text, markup language, tokenization, and character encoding.

The initial annotation of the LDC Gigaword corpora (in English, Chinese, as well as French) has been used for various computational linguistics research projects, and had as an initial goal to help with statistical machine translation and parallel corpora creation. They were also used to create a parallel (bilingual) corpus for the French/English language pair of a statistical machine translation system with information retrieval techniques (Abdul-Rauf & Schwenk, 2009). The literature search undertaken for this study has also led to a series of other published articles using the Gigaword corpora. (See Appendix A).

The FGC, a critical mass of vocabulary used in written and spoken French, offers another valuable source of comparison to evaluate untested CEFR vocabulary selections against objective usage-based frequency data. It is important to note that textual sources of this study were not completely mutually exclusive. The FDF included newswire also contained in the FGC. Textual overlap of this study's two primary corpora sources represented, however, less than

0.0001% (1/2267), according to Lonsdale (2012), and it was hoped that word overlap between the FDF and the FGC would be close to, if not 100%.

Important Considerations Relating to Vocabulary Selection

This section shows how the rate, quality, and quantity of lexical content used to instruct language learners has been evaluated. This evaluation is important to continue to refine the lexical content accounting process. Indeed the literature did not give precise answers on the number of words and the words needed to function at a given language proficiency level. The literature also did not explain the best method to count lexical units for instruction or the best use of corpora to answer the study questions. However, these questions sound reasonable and should be measurable and quantifiable. Answers to these questions with older and less sizable data sources have been shown to be valid language performance indicators. These answers seemed, however, hard to obtain and the little research done so far in this respect showed that they vary from one language or one level to another. The difficulty arised partly from a lack of coherence of proficiency standards across languages and no European mandate and finances from national governments (Trim, 2011). Lexical studies are important but not as high as they might need to be on national research agendas. However, despite a shortage of lexical research data, teachers are known to use assessment involving vocabulary the most frequently of all (Brumen, Cagran, & Rixon, 2009).

CEFR French profiles could give a more explicit measure of rate, quality and quantity. It could delimit vocabulary size but it introduced the following paradox. On one hand, European foreign language experts did not want to impose a standardized approach to language learning, but, on the other, they wanted to harmonize and come to a vast comparative assessment of European language learners. To produce these profiles, use was made of the communicative

notional task-based approach but no number or word specification was attached to them, thus leaving profile user wondering why they should teach profile vocabulary over any other vocabulary selection they might choose.

Theoretical underpinnings. An explanation of the principles underlying the communicative task-based approach and directing French profile selections was presented earlier. The communicative approach, as its name suggests, is not considered a theory. This section explores further how theory (or the lack thereof) influences what words, how many words, and the rate of words learned in FFL learning and teaching.

If research has done anything at all in the area of vocabulary studies, it is to stress the reality that "CEFR scales [are] lacking empirical support of what L2 specific knowledge and skill is minimally required for performance considered adequate in terms of communicative functioning" (Hulstijn, 2007). The scaling of available CEFR descriptors was done in the absence of fully developed and properly tested theories of language proficiency (Hulstijn, 2007 quoting North and Schneider, 1998). Table 9 presents vocabulary criteria at play and the learning outcomes areas they influence.

Another reason for this study related to what is known about beginning and intermediary level definitions. They have been loosely defined and need a narrower definition. Loose level definitions allow for wide interpretations, and thus variations in lexical selections, lexical sizes, lexical coherence, lexical progression, and ultimately lexical competence. These definitions also influence pedagogy at the curriculum design, implementation, and testing stages. The findings of the present study was to define vocabulary dimensions with more precise specifications and help researchers fill Table 10 with empirical quantitative and qualitative data.

Table 9

Critical Factors of L2 Instructional Design Related to Vocabulary

Vocabulary	How many words?	What words?	Word rate?
Criteria	Size	Specification	Growth
Learning Outcomes	Word quantity Breadth tasks, domains, functions, notions, situations, locations, topics, and roles	Word quality Depth precision in the use of language, understanding of meaning, expression of meaning	Progression over time, core proficiency, advanced proficiency

Table 10

Vocabulary Dimensions by Proficiency Levels

	CEFR Levels			
Dimensions	A1	A2	B1	B2
Size	vocabulary size?	vocabulary size?	vocabulary size?	vocabulary size?
Frequency	What and how many frequent words?	What and how many frequent words?	What and how many frequent words?	What and how many frequent words?
Progression	A1 word base	A2 additions to word base?	B1 additions to word base?	B2 additions to word base?

Definition of comparison standards. Other research established that fifty percent of students fell below the B2 level of competence in English and had an inadequate knowledge of Nation's academic word list. These findings indicated that neither level B2 nor knowledge of the academic word list were valid as an entry level for tertiary education. They were corroborated by similar studies at other universities and made it doubtful whether level B2 can be reached by high school graduates (Platzer, 2006).

With such discrepancies, no one group of foreign language learners is comparable to another. The CEFR, initially intended as a "framework" for comparisons could not become a

standard of comparisons without first establishing comparable levels, and within these levels comparable functional vocabulary in size and kind. Only an agreed-upon vocabulary base for each level would allow European communities to meet the requirement of aligning curriculum and assessment to the CEFR scales, and be accountable for learning outcomes (Fulcher, 2008).

Moreover, no convincing research had been conducted to establish the empirical soundness of CEFR levels. Hulstijn (2007) in his review stated that no longitudinal studies establish that

> All L2 learners at some functional level other than A1 (e.g. B2), arrived at that level by passing the level below (B1, in this example). In other words, there is no empirical evidence that, in overall oral proficiency (CEFR, 2001, p. 58) for instance, all learners first attained the functional level of A1, then the level of A2, etc., until they reach their highest level.
>
> [No empirical evidence shows that] all L2 learners at a given level (other than the lowest level A1), are able to perform all tasks associated with lower levels which should be the case if the CEFR scales are genuinely implicational and unidimensional"... and more seriously, there is no evidence in terms of learner performance that a learner at a given level of an overall scale (e.g. B2 overall oral production) necessarily possesses the quality in terms of the linguistic scales at the same level for the other dimensions of the level (e.g. B2 vocabulary range, B2 grammatical accuracy, and B2 phonological control). (p. 666)

It is also important to note that sequencing of content by level became a crucial component of design only if the instructional approach chosen was focused on content-driven language

learning (Kumaravadivelu, 1993). The CEFR vocabulary profiles were more focused on tasks and notions.

Thus, Hulstijn (2007) concurred with Fulcher and reiterated the challenge of developing and testing theories of language proficiency, asking for research efforts on linking developmental routes and second language acquisition with language assessment, and suggesting the importance of corpus research.

> On the basis of statistical analysis conducted on corpora of a wide variety of oral and written discourse, it should be possible to define the set of words and multiple word constructions that have a high probability of occurring in certain communicative situations, especially in situations that all adults are likely to be able to deal with. It is high time that researchers of second language acquisition, researchers of language assessment, and corpus linguists paid attention to each other's work and engage in collaborative research, testing the linguistic, psycholinguistic and sociolinguistic assumptions on which the CEFR rests.
>
> To date, the CEFR only rests on teacher perceptions. Valid and reliable as they are or may be, they provide a foundation too weak for the CEFR building with its heavy-weight implications for language education policy in Europe. (p. 667)

Hulstijn's theoretical premises of language proficiency involving (a) a global distinction between lower-order and higher-order cognition in language processing, (b) universal human capability for implicit learning, and (c) the notion of core language proficiency were studied at the University of Amsterdam Center for Language and Communication (2007-2011). Regarding his research, Hulstijn suggested that the limit between core language proficiency and peripheral or advanced language proficiency "be seen in probabilistic terms" (2007, p. 664).

Hulstijn's critique (2007) of the CEFR's lack of theoretical grounding also outlined a whole research agenda to which this study hoped to contribute. As Hulstijn argued, a definition and description of language proficiency by CEFR-levels could not be undertaken unless matched to quantity (the number of proficiency factors a user masters, vocabulary size being one of these factors) and quality (degree of effectiveness, efficiency, or precision), where one might observe a learner being on a B1 quantity level, but on a C1 or C2 level in terms of quality – or any other learning outcome scenario.

To put vocabulary growth in the context of CEFR levels, Milton (2006) presented the figures of vocabulary size tests for French as a foreign language in Britain and English as a foreign language in Greece and Hungary (See Table 11). An exact correspondence between the English and the French language was not expected. However, French as a second language was at the lowest end of the CEFR vocabulary standard. The CEFR aimed at a more ambitious scale of learning for students of any foreign language, not only French. The inventory of the CEFR French vocabulary profiles yielded a count of 6,486 lexical units for level B2 instruction input (non-lemmatized) and this amount represented the minimum lexical input a student could learn. Table 12 shows the cumulative amounts by level inventoried from the CEFR French profiles.

Given the lack of research published on vocabulary growth and size from instruction of French as a foreign language, few tried to determine how the vocabulary of students of French grows over time. However, vocabulary size is the best predictor of success at major British languages exams (Milton, 2006).

Table 11

Number of Known Lemmas at Various CEFR Levels in Britain, Greece, and Hungary Based on Exam Scores

CEFR level	Wordlist Size	French in Britain	English in Greece	English in Hungary
A1				
A2	1,000	850	2,000	
B1	2,000	850	3,000	3,100
B2		1,920	3,500	3,900
C1				
C2		3,300	4,500	

* Excerpted from Milton (2006)

Table 12

Description of the CEFR French Vocabulary Profiles

Vocabulary profiles characteristics	Niveau A1	Niveau A2	Niveau Seuil	Niveau B1	Niveau B2
Date published	2007	2008	1976	2011	2004
Authors	Beacco & Porquier	Beacco et al.	Coste et al.	Beacco et al.	Beacco et al.
Quantity of lexical data (lemmas, types, other units)	1,525	2,377	3,997	3,670	6,486

Vocabulary parameters. The literature did not clarify important vocabulary parameters necessary to identify and select vocabulary content for curricular design. Lexical growth rate, quality, quantity, size and frequency, and accounting were, of course discussed but no definite consensus was discerned.

Growth rate. Learning vocabulary has been qualified as "the core component of all the language skills" (Long and Richards, 2007, p.xii). Without this kind of learning, nothing else in a language can happen. Moreover, despite the fact that the CEFR relied on an important construct labeled language proficiency and provided scales of linguistic competences such as vocabulary

range and vocabulary control, little theoretical grounding has been undertaken to establish the "quantitative and qualitative dimensions of language proficiency" (Hulstijn, 2007), and, so far, no smooth cumulative language learning progression has been observed. Hulstijn has demonstrated that without tested theoretical grounding for language proficiency and explicit quantitative and qualitative content specifications linked to CEFR levels the task of test developers remain shaky (cf. Fulcher & Davidson, 2007). Trim (2011), a CEFR veteran expert, also recognized the unevenness of the concept of CEFR level since natural breaks in the growth process of language learning are hard to find, and said that the representation of progress in learning as a succession of discrete levels arises not from the nature of language learning itself but from the necessities of the social organization of learning.

Milton (2010) in his review of communicative proficiency and linguistic development also pointed out two important truths, namely that

> . . . progress through the [CEFR] hierarchy is closely related to vocabulary knowledge and knowing more and more words in the foreign language. High level performers tend to have extensive vocabulary knowledge and elementary level performers do not. The second is that knowledge of the most frequent words in the foreign language appears crucial to successful performance. (p. 218)

Thus, the description of comparable language proficiency levels was an effort to set evaluation standards based on loosely defined tasks a learner can accomplish, not on existing language proficiency theory. To specify and normalize proficiency standards, countable vocabulary input would be necessary. This was why the untested and uncounted representation of progress gave an enormous challenge to developers who want, for instance, to compare L2 learners and to connect their test results to each different CEFR level since learners' response to task-based

statements will vary greatly in quantity and quality (with no existing definition for minimally adequate responses). These responses would be strongly connected to vocabulary input presented and learned but as Fulcher (2004) explained CEFR scales are not based on tested theory and not linked to content specifications; thus they could not reasonably provide equivalence and comparability (Alderson, 2007; Fulcher & Davidson, 2007; Weir, 2005).

The six reference levels (Table 1) are becoming widely accepted as the European standard for grading an individual's language proficiency. They are also recognized outside of Europe even if some institutions have kept their own naming conventions, e.g. "intermediate". The CEFR levels have been popular because they were successful at using the already familiar labels: beginning, intermediate, advanced, with the new A1, A2, B1, B2, C1, and C2 level distinctions. Language and certification programs evaluate their own equivalences against the CEFR. Levels, similar to those used in the CEFR, are used in North America, for example with the American Council for the Teaching of Foreign Languages (ACTFL), but so far, they remain without ties to specific corpora or vocabulary inventories.

Table 13 shows that the ILR, ACTFL, and CEFR level boundaries are unclear, and elicits several questions. What words and numbers of words might be attached to these levels? What do proficiency guidelines say about the number of words a student should master? They do address the "kinds of words" issue by specifying, for instance at A1 level: isolated words and phrases, concrete, everyday, very basic, question words, but they do not even give a word number range by level. The approach is minimalist but the CEFR does not indicate a minimal number of words that should be accounted for at each level.

Table 13
Corresponding Levels of Linguistic Proficiencies by Classification System

ILR	ACTFL	CEFR
0 / 0+	Novice – Low, Mid, High	A1
1	Intermediate –Low & Mid	A2
1+	Intermediate – High	B1
2	Advanced – Low & Mid	B2
2+	Advanced – High	B2
3 / 3+	Superior	C1
4 / 4+	Distinguished	C2
5	Native	

The European and North American scales do not have equivalent levels however. When the CEFR proficiency scale is compared with the Interagency Language Roundtable Scale (ILR, United States) and the ACTFL, the correspondence is hard to establish because it does not agree with the generally accepted scale where Novice, Intermediate, Advanced and Superior would correspond to 0/0+, 1/1+, 2/2+ and 3/3+, respectively on the ILR scale. In a panel discussion at the Osaka University of Foreign Studies, one of the coauthors of the CEFR, Brian North, stated that a "sensible hypothesis" would be for C2 to correspond to "Distinguished," C1 to "Superior," B2 to "Advanced-mid," and B1 to "Intermediate-high" in the ACTFL system. Moreover, the characteristics of a pre-A1 proficiency level were not described in the CEFR. The assumption is that it would correspond to no knowledge of the foreign language to be learned at all. It was further assumed that lexical competence, or vocabulary knowledge, at the beginning of one level is measured by the mastery of vocabulary encompassed at all levels below, i.e. C1 advanced vocabulary will include all A1, A2, B1, and B2 lexical items. The example of Rolland and Picoche (2008) who proposed to systematically learn the "current usage, basic" French lexicon at level A1 and beyond was a case in point. The authors illustrated the existing level boundary

dilemma as they listed 3,357 frequent word units but decided not to disambiguate CEFR levels of progression and proficiency.

Quality. Vocabulary knowledge is important to learners' perception of linguistic competence (Kelly, Li, Vanparys, & Zimmer, 1996). It boosts learners' speaking, listening, self-confidence (Harlow & Muyskens, 1994). With vocabulary development, self-perception could evolve positively over time and reinforce L2 fluency and vocabulary acquisition (de Saint Leger, 2009). Vocabulary quality was associated with what characterized knowledge of words themselves, such as their forms, meanings, difficulty, and kind. Vocabulary growth and rate, as mentioned earlier, depended on "word knowledge", a construct reviewed by Gardner (2007). The term *word* implies single or multiple orthographic elements sequenced in a way to produce meaning through lexical and semantic processes such as fossilization, and word-formation rather than through imposed rules. Nation (1990) characterized word knowledge as knowing (a) the degree of probability of encountering the word in speech or print; (b) the limitations imposed on the use of the word according to function and situation; (c) the syntactic behavior associated with the word; (d) the underlying form of a word and the derivations that can be made of it; (e) the associations between the word and other words in the language; (f) the semantic value of the word; and (g) many of the different meanings associated with the word. Word knowledge is, thus, strongly associated to word frequency.

Depth corresponds to "knowledge of specific words or the degrees of such knowledge" (Wesche & Paribakht, 1996, p. 13). Depth of knowledge was an indicator of lexical progression. For instance, advanced learners needed depth and speed of access as well as range in their vocabulary knowledge for ease, precision, and effectiveness in their linguistic expression (Wesche & Paribakht, 1996).

Word knowledge was strongly associated to L2 reading competence. For instance, Qian (2008) was able to relate word knowledge to reading competence on the TOEFL exam, and concluded that, in assessing reading performance, "discrete-point vocabulary items and fully contextualized vocabulary items provide a similar amount of prediction" (p. 2). It was also associated to listening comprehension. L2 reading and listening comprehension are the first two receptive stages of language acquisition. Attention to improving L2 proficiency in these two areas should be essential in order to improve productive stages of language acquisition (speaking and writing). For all four stages L2 vocabulary acquisition was essential and should be emphasized first. In a study of vocabulary in writing assessment with a focus on the lexical frequency profile, attempts to ascertain the percentage of words a writer uses at different vocabulary frequency levels were made. It was assumed that the more proficient writers use more words of lower frequency but results were difficult to compare (Laufer & Nation, 1995).

Studies have also been conducted to investigate the relationship between L2 vocabulary knowledge and success in reading comprehension, and subsequently, to find the vocabulary threshold. The vocabulary threshold is the minimal vocabulary that is necessary for “adequate” reading comprehension. Information on lexical threshold was important for second language education, particularly for courses with reading as their main focus, since such information may help teachers and course designers in setting vocabulary goals and designing lexical syllabi. Thus, for example, if the lexical threshold was found to be 7,000 word families, then by the end of a course in academic reading, students should try to reach this vocabulary size if they intended to engage in reading authentic academic material (Laufer & Ravenhorst-Kalovski, 2010).

Vocabulary is the main building block of language and is a necessary component of language learners’ development. Vocabulary learning was linked to word frequency for all

learners at all stages (David, 2008). No evidence suggested that the ability to acquire vocabulary became inoperative at any age (Singleton, 1998). Word knowledge was closely related to word recognition and spelling in L1 lexical acquisition (Lete, Peereman, & Fayol, 2008). It allowed implementation of principles of universal grammar to be available from the outset of language learning (Jacubowicz, 1989). Mastery of core lexis was an essential component of reading proficiency (Upjohn, 1999). Cummin's interdependence threshold hypotheses assert that L1 knowledge could transfer, but only after learners attained a threshold of L2 knowledge (Brisbois, 1995; Cummin, 1981). It was also observed that the acquisition of less common vocabulary made a major contribution to students' progress during their year-12 course and it was an important factor in individual differences in overall achievement (Richards, Malvern, & Graham, 2008). However, even though vocabulary is necessary to a learner's development, an ample vocabulary and mastery of grammar were not sufficient for the understanding of a foreign-language text. Knowledge of the cultural practices governing the use of grammatical and lexical rules was indispensable (Beacco, 1981). Moreover, researchers noted that there was a relationship between the difficulty of a word and lexical acquisition order (Vermeer, 2004). Genus terms or prototypes like *bird* or *chair* were more frequently heard and used in everyday speech than super categories or subcategories, and so were learned at an earlier stage (Van der Vliet, 1997).

Detailed knowledge of words and their attributes were also assessed through "breadth" tests. Much has been written about this concept (Cohen, 1986; Gass, 1989; Nation, 1990; Read, 1989; Richards, 1976; Robinson, 1989; Wesche & Paribakht, 1996). This study tried to identify what frequent French words would be important to know based on their usage as well as their relevance for the completion of communicative tasks.

Quantity. European governments want to make foreign language learning assessment comparable across languages. To accomplish this complex task, the native lexicon, often captured in comprehensive monolingual dictionaries, is a key component and shows the value of vocabulary size. Laufer also pointed out the value of being able to assess the progress of L2 learners' vocabulary size for language research and pedagogy, i.e. knowing how much instruction would be needed, and how lexical syllabi could be more realistically planned to reach a vocabulary threshold level necessary for the comprehension of written authentic prose (Laufer, 1998). Quantified vocabulary teaching and testing helped students become aware of the size of their lexicon (Mondria, 2006; Shillaw, 1995). A more precise idea, though not exact, on word quantity was possible thanks to existing research based on published CEFR-related vocabulary profiles, exam estimates, and coverage studies.

At this time considerable discrepancies exist between counts of vocabulary sizes of European languages, and counts of vocabulary sizes at each progressive level within a language. These discrepancies destabilize the system of equivalences the European Council wants to establish with the CEFR and the European Survey on Language Competences (ESLC).
Table 14 reveals the wide word-number ranges, thus far identified for CEFR levels that have been mentioned in publications or counted in language word profiles (Decoo & Kusseling, 2009). To understand Table 14 one has to know that, for instance, the total of 6,800 entries in the French B2 Profile (Beacco et al., 2004) represents 6,214 indexed lexical items and grammar items, where references to functions and notions in the index were not counted, and homonymous and strong polysemous words were sorted out. Table 14 also points out the heterogeneity of data sources and lexical characteristics which hamper comparability.

Table 14

Word Count Estimates Identified for CEFR Levels

				CEFR Proficiency Levels					
Language	Authors	Data Source	Lexical Characteristics	A1	A2	B1	B2	C1	C2
English	Van Ek & Alexander 1980	Profiles	lemmas, types, multi-word units		700	1,100-1,500			
English	Van Ek 1976	Profiles	lemmas, types, multi-word units			1,600			
English	Meara & Milton 2003	test scores	lemmas	<1,500	1,500-2,500	2,750-3,250	3,250-3,750	3,750-4,500	4,500-5,000
English	Schmitt 2008, see also Nation 2006	estimates based on CEFR descriptors and coverage criteria	word families						15,000*
English	Bergan 2001	Estimates and profiles	lemmas		850	1,500	4,500		
French	Coste et al. 1976	profiles	lemmas			3,000			
French	Beacco et al, 2004	profiles	lemmas, types, multi-word units	1,000	1,700	(4,000)	6,800		
French	Rolland & Picoche 2008	Frequency Dictionary	lemmas	3,357					
French	Milton 2006	test scores	lemmas	(400)	800-1,000	800-1,000	2,000		3,300
Spanish	Instituto Cervantes 2006	Profiles	lemmas, types, multi-word units	1,300	3,000	7,000	14,000	21,000	30,000

* Counting by word families has been transposed into words (mainly lemmas), using the 1.7 ratio. Figures in parentheses indicate estimates based on proximate level figures.

It has already been shown that vocabulary input elicited from CEFR descriptions has a wide range within each level from one language to another (Kusseling & Decoo, 2009). Table 15 exemplifies how CEFR vocabulary description influenced this range within a level and along the whole scale. Descriptor imprecision, e.g., the use of the terms: *basic*, *sufficient*, *good range*, *very broad*, continues up to level C2 (CEFR, p. 112). This imprecision allowed for wide discrepancies in vocabulary range and size. However, the CEFR stated that "size, range and control of vocabulary were major parameters of language acquisition and hence for the assessment of a learner's language proficiency and for the planning of language learning and teaching" (p. 150).

Table 15

Vocabulary Descriptions and Word Number Ranges Counted for CEFR Levels

Level	Vocabulary Description	Word Number Range	Difference
A1	***basic*** *vocabulary repertoire*	400 – 3,357	89%
A2	***sufficient*** *vocabulary*	700 – 3,000	77%
B1	***sufficient*** *vocabulary*	1,100 – 7,000	84%
B2	***good range*** *of vocabulary*	2,000 – 14,000	86%
C1	***good command*** *of a* ***broad*** *lexical repertoire*	3,750 – 21,000	82%
C2	***good command*** *of a* ***very broad*** *lexical repertoire*	3,300 – 30,000	89%

Profile word count. Since, even with the caveat that vocabulary profiles did not decide what learners had to learn at a given level but that they were free to choose their vocabulary based on needs, panel experts, instructional designers, and test developers still hoped learners would use them. An important common feature of a language proficiency level for these stakeholders was the profile word count. One method of calculating vocabulary quantity to be learned consisted in estimating the size of the learner's lexicon. That size might be considerable,

adequate, or too small for comprehension or expression at a beginning, intermediate, or advanced level. Vocabulary size tests are the current popular method to estimate the number of words a learner knows. Vocabulary tests related to curricula-dependent assessments and placement are only approximate indicator of learners' vocabulary size however. Currently, DIALANG, an internet-delivered computer-adaptive diagnostic test for 14 European languages is based on the CEFR diagnoses, among other competence areas, vocabulary and vocabulary size. It can be used in the context of language learning, for progress or achievement tests or to diagnose learners' strengths and weaknesses. Adaptation to the learner's response can be based on vocabulary item difficulty or item content (language features such as overall frequency of words in the language).

Researchers have estimated that most native speakers have a vocabulary size in the range of 17,000 word families (Goulden, Nation, & Read, 1990). A word family is related by a common base form, to which different prefixes and suffixes are added. There can be inflected or derived forms (e.g. climb, climbs, climbing, climbed, climber, climber's, climbers'). Estimates of second language learners' lexicon were calculated. They varied depending on the definition of the concept word knowledge, linguistic program goals, didactic approaches, needs and skill levels of learners. For instance, intermediate learners was estimated to need knowledge of some 5,000 word families but certain test specifications only list some 1,600 words (Maun, 2009). In the mid-1990s, students' total English vocabulary was estimated for the first time by use of the computerized Eurocentres Vocabulary Size Test, administered upon students' entrance at the university and six months later. Overall results indicated important vocabulary growth, but students with high proficiency levels did not increase their vocabulary size significantly. When administered for French in a British school and university, the test of vocabulary size chosen showed that only by the end of the third year of university did vocabulary growth reach 3,300

lemmas, as determined by students' average test scores (Milton, 2006). According to Milton's research, students of French in the British system learned less French then they learned 50 years earlier, 3-4 words per study hour which ended up being an average of 2,000 lemmas by the end of seven years of French studies as shown in Figure 2.

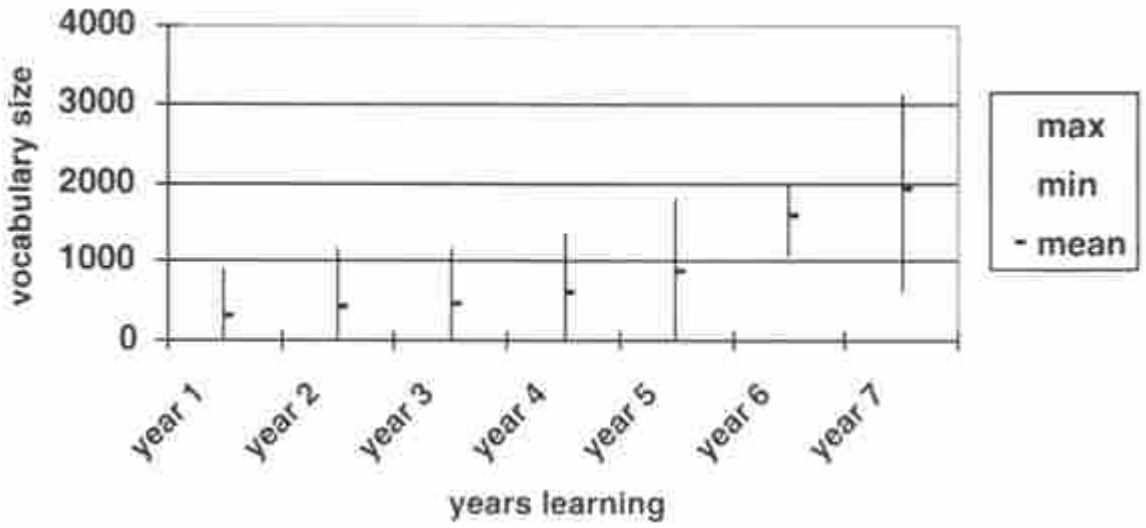

Figure 2. Milton's diagram for progress in vocabulary size by school year

A problem with vocabulary size estimates is, however, that there are no reliable tests of vocabulary size. Nation's Vocabulary Levels Test (1990) is probably the test that is closest to a standard test of vocabulary size. The main reason for unreliability is sampling, i.e. most sampling methods are biased in such a way that they make it more likely for common words to appear in an apparently random sample. The tendency of vocabulary size tests is overestimation of vocabulary size on tests that are too short, with items that are not normalized on large populations of testees relative to a predetermined norm and do not differentiate between breadth and depth of vocabulary knowledge (Read, 1993; Wesche & Paribakht, 1996), and with no precise idea of the number of words that make up the vocabulary targeted, which is the critical variable for constructing a test of vocabulary size.

A solution to the deficiencies of tests of vocabulary size has been checklist tests. Checklist tests make use of a set of real words and a set of imaginary, non-existent words (Meara & Jones, 1988; Meara & Jones 1990; Meara, 1990). Another solution is the use of statistical techniques based on "Signal Detection Theory" (McNichol, 1972). Yet, another way to remedy deficiencies of vocabulary size tests consists in using a battery of tests of different frequency bands or of different specialized areas of lexis to build up a profile of a testee's vocabulary knowledge and to measure vocabulary growth over relatively short periods of time. These tests have moderate correlation with tests of other linguistic skills and with other vocabulary tests not attempting to measure vocabulary size. Testees have an overtendency to say *yes* to imaginary words. Frequency band tests do not work well with low-level learners, and with French or English native speakers, for instance, the close relationship between the lexicons of English and French might mean that vocabulary size per se is less important for learners of these two languages than it is for speakers of Japanese or German. These frequency band tests appear to be reliable enough to allow a look at vocabulary growth and factors affecting that growth at advanced levels.

A second way of assessing vocabulary quantity has been to determine coverage (defined in Chapter 1). This includes the number of words derived from frequency lists needed to read a particular text, or better a particular representative corpus with varying degrees of ease. Indeed, vocabulary size was strongly connected to coverage and is a good indicator of general knowledge and language proficiency.

> In both English and French the most frequent 2,000 words, and overwhelmingly the most frequent words in a language are learned earliest and give about 80% coverage of normal text. This is a very interesting and important figure because it marks the level at which

> learners appear to progress from understanding almost nothing they hear or read, except in the most limited and contrived of circumstances, to having passages of clarity and being able to grasp the gist of a conversation or a reading passage. But, in both languages, to add sufficient vocabulary to understand the remaining 20%, and therefore understand all of the text, requires massively more vocabulary. Learners do not have anything like full comprehension of a text until they have at least 95% or 98% of a text and that may require 6,000 to 8,000 words. (Milton, 2006, p. 3)

Size and frequency. Francis and Kučera (1982) showed the effect of vocabulary size on language comprehension in their study of English texts totaling one million words (see Table 16). They found that learning the most frequent words in an English text provides a comprehension of most of the words in those texts.

Table 16

Effects of Vocabulary Size on English Language Comprehension

Vocabulary Size (Number of Lemmas)	Written Text Coverage
1,000	72.0 %
2,000	79.7 %
3,000	84.0 %
4,000	86.8 %
5,000	88.7 %
6,000	89.9 %
15,851	97.8 %

The figures seemed to improve for coverage of words learners heard in informal speech. Using a 5 million word spoken corpus compared to the Schonell et al. (1956) 512,000 word corpus, Adolphs and Schmitt (2003, 2004) found that 2,000 word families supplied lexical coverage for less than 95% of spoken discourse based on the CANCODE (Cambridge and Nottingham Corpus

of Discourse in English). The research numbers reported here were derived from lemmas. The same amount of word families would give an even higher coverage. However, one should not forget that the number of words cited for one language may differ considerably for another.

Laufer, in a 1992 study of general vocabulary based on frequency lists only, suggested that for general reading comprehension "the minimal number of words constituting the lexical threshold" (p. 129), or minimal turning point of vocabulary size for reading comprehension, was 3,000 word families. So, unless L2 readers had reached this lexical level they would be "hampered by an insufficient knowledge of vocabulary" (p. 130), even if they relied on technical vocabulary to compensate for general vocabulary below 3,000 word families:

While it was true that the knowledge of technical vocabulary was helpful, its value should not be overestimated. Empirical evidence showed that it was the general -- not the technical vocabulary -- that was most problematic for learners (Cohen, Glasman, Rosenbaum-Cohen, Ferrara, & Fine, 1979). Baudot's French frequency lists (1992) had been used for vocabulary size tests. His most frequent 5,000 words provided however less than 95% coverage of his 1.2 million word corpus.

Research also showed that vocabulary expansion based on word frequency accelerated the reaching of better coverage (Bogaards, 1994; Hirsh and Nation, 1992). The most frequent words in a language provided most coverage of a text and were very useful to prepare for advanced studies of a language. This relation between frequency and coverage remained steady up to 11,000 words learned (Hazenberg & Hulstijn, 1992). Research further showed that 95% to 98% constitute minimum coverage for fluid reading (Laufer, 1989; Nation, 2006). Thus, teaching materials introducing the more frequent words first (in layers of e.g., 1,000 words) not only is helping learners reach higher coverage sooner, but also avoid the learning of unimportant words.

In reality, CEFR profile counts leading to a more precise indication of minimal vocabulary size could not actually be compared since their unit of analysis varied from source to source. Not accounting for vocabulary size would be a way of making the educational system fail in its goal of ensuring social, cultural, and economic integration and development. Since narrow range of vocabulary (passive and active) has been associated early on with the second most important weakness in language learning (Furness, 1975), and is a vital cause for lack of employability and mobility (Zavasnik, 2009).

Word accounting. To continue the discussion relative to answers not yet yielded by this study's primary sources, it could be said that no matter the degree of importance granted to notions and tasks in communicative language use, words remain the most usable units to count and measure vocabulary growth, content, and size. Bauer & Nation (1993), Cowie (1992), Hazenberg & Hulstijn (1996), Meara (1996), and Gardner (2007) discussed methods for vocabulary counting. Unclear or not-agreed upon definitions of lexical units of measurement have produced confusing and diverging results. Indeed, they have given a skewed idea of the number of words a foreign language learner acquires in order to improve proficiency or increase coverage and fluency. To date, there is no consensus on number and kinds of words monitored in foreign language teaching that would allow comparisons. The lack of consensus originated in great part from a divergence on focus. Some focused on word form or word meaning or both. The question arose on how this focus was pertinent to instructional design. When focusing on word meanings, *word meaning*, *lexemes* or *didactemes* might be used as units of analysis and counting.

To count *word meanings*, each meaning of a word unit would be counted separately (This would include meaning of homonyms and homographs which are much more present in the

French language than in the English, such as these examples: *louer* (to rent and to praise), *joue* (joue= play, la joue=the cheek)).

Decoo has also referred to the concept of *didacteme* to count vocabulary input, which he defined as "any minimal quantifiable unit dependent of language level and capabilities of short-term memory that can be identified in a certain context and that a learner receives as input in a single mental movement" (2011, p. 20), adding that "a certain amount of precisely defined didactemes from several hundreds to several thousands will make up the content of specific units or levels." (2011, p.22). Examples of didactemes are *avoir faim* (to be hungry) or *avoir sommeil* (to be sleepy) where the French use the auxiliary verb *avoir* when the English use the auxiliary verb *to be* to express the same concept.

Today the more popular unit of measurement would be the *lexeme*. With this concept what counts as one *lexeme* is any lexical unit with a single meaning, often a multi-word unit regardless of the number of words it contains. The literature showed that learning new meanings of already known wordforms would add a level of complexity to vocabulary size counting (Bogaards, 2001). Methods used for analysis of word meaning are complex and time consuming even with state-of-the-art software technology able to program deep parsing and semantic analysis. For instance, the main problem with identifying and counting words and frequencies in electronic corpora has been the difficulty of counting semantically-based *lexemes* vs. form-based *types*. For example, out of 600 form-based *types* one could generate 3,000 semantically-based *lexemes*.

Now, when focusing on word forms which was the scope of this study, researchers might use *word families*, *lemmas*, or *types* as units of analysis. When counting by *lemmas***,** the word base and all its inflected forms would only represent one count. For instance, only the lemma

work would be counted and listed as the representative for *work, works, worked, working, wrought*, whether these word forms have one or more meanings would be irrelevant for the counting.

The identification of dictionary *lemmas* could tell more about *word families* since both of these categorizing concepts rely on a base form shared by a group of words. It is important to consider what happens when the morphological counting unit changes. For instance, Bauer and Nation (1993) calculated that 800 *word families* equal 1,400 words. Were words counted by *word families*, for instance, a *word family* unit would include the base word plus all of its inflections and its common, transparent derivations. This counting method was illustrated by an example taken from Laufer on lexis needed for reading comprehension: "the knowledge of *observe,....*subsumes the knowledge of *observation, observable, observant, observance* and all their inflections. ...3,000 *word families* represented in terms of dictionary lemmas, would be 3,000 x 1.6 = 4,800 (1992, pp. 129-130). In conversions from word families to lemmas to types, the inflectional system of a language, e.g. French or German compared to English, needs to be taken into account since it will considerably influence the conversion ratios obtained. French profile authors were not concerned about counting words and associating numbers of words to language proficiency levels (Milton, 2006). They inventoried lemmas but not lemmas exclusively. Multi-word expressions, plural or feminine forms were found, the reality being that the vocabulary profiles would include infrequent inflexions and derivations, and the assumption being that all inflections not listed in the vocabulary profiles would somehow be recognized and learned.

Counting in *lemmas* or *word families* might facilitate language comparisons. It would not, however, give a full picture of language attainments. Thus, counting *types*, as defined in Chapter

1, reflected both the inflectional and derivational systems of a language and would allow a more precise count of learners' vocabulary size and growth. Methods of morphological analysis such as shallow parsing were, at this time, the more manageable way to account quantitatively for existing wordforms.

Use of corpora for lexical selection. Counting words and word accounting automatically pointed to corpus studies which allow the determination of word usage in comparison to what can be accomplished with these words, i.e., the task-based approach. The study of language through corpus-based linguistics research distinguishes itself from other methods by its systematic observation and analysis of real usage-based language evidence (the corpus itself). The use of specifically designed corpora allows for a more objective scientific analysis of word usage. It is an antidote to the criticism addressed to instructional designers, using the notional syllabus model, for lacking "empirical evidence upon which to base their selection of structures and exponents when working within a functional framework, and (. . .) [there being] an unsatisfactory reliance on intuition" (White, 1988, p. 82)

As discussed earlier, the CEFR vocabulary profiles are not based on quantitative usage studies but rather on introspection and experience. Little explanation was given regarding the procedure the authors followed for creating their vocabulary inventories. As Schonenberg (1988) observed in her work the profiles are characterized by *a priori, general* and *minimal* features:

> *"a priori":* it is not the result of a rigorous L2 needs assessment; sociolinguistic surveys were not conducted which could have revealed specific language use in given situations. To the contrary, its authors trusted their intuition regarding language segmentation.
>
> "*general*": the authors' intent was to interest a majority of L2 learners. Thus their work does not respond to highly specialized linguistic needs.

"*minimal*": with minimal communicative competence level (threshold level) learners when expressing themselves in their L2 should be able to be easily understood by natives; they can achieve this goal by speaking correctly enough and with an acceptable flow; moreover, L2 learners should be able to understand most of what native speakers say without them having to try hard to make themselves understood. (p. 12)

The French profiles are unlike their English sister publications which have been described as part of a long-term extensive research program using the Cambridge Learner Corpus (CLC), a growing collection of several hundred thousand examination scripts written by learners from all over the world, combined with solid evidence of use in many other sources related to general English, such as examination vocabulary lists and classroom materials to confirm what learners can and cannot do at each level, and informed by the Cambridge English Corpus, a multi-billion word corpus of spoken and written current English, covering British, American and other varieties, and reflecting what learners do know, not what they should know (English Profile, 2011).

Corpus linguistics has been used to evaluate the content of instructed language teaching and learning. These evaluations are guided by frequency, an important principle for language proficiency, Also, comparisons have been made to the larger lexicon and usage-based frequencies (see Davies & Face, 2006).

Word frequency has accounted for both vocabulary quality and quantity. Research has shown its role in vocabulary learning, e.g., the cost benefit of learning frequent, infrequent and specialized words (Nation, 2001). Already observed in L1 with very young learners high frequency words are better known than low frequency ones in children independent of text length and syntactic ability (Brown, 1993; Kibby, 1977; Nation, 1990; Vermeer, 2001). The relation between word

frequency and acquisition order is more complex for adults because adults know more words and after the most frequent 12,000 words, enormous numbers of words have approximately the same frequency (Hazenberg, 1994). The relation of most frequent words learned first does not hold for many academic words, in particular for L2 learners, because of their interlingual roots and international use. High frequency function words such as prepositions and conjunctions which have little semantic meaning or high frequency general content words such as *personne* (ranking 84 in the FDF and meaning *person*, *people*, *anybody*, *anyone*, or *noboby*). Corpus linguistics frequency studies can identify more easily highly frequent words and gave empirical evidence that showed they are the most difficult to learn (Cohen et al., 1979; Laufer, 1992). However frequent words are more useful to students receptively and in production, whereas relatively rare words prove less useful in the earlier stages of language learning (Biber & Reppen, 2002). Vocabulary learning was linked to word frequency for all learners at all stages (David, 2008).

In the past, vocabulary experts have relied more on their own observations and intuitions, since building mega-corpora was a dream, but today, opportunities to base lexical selection guidelines on researched word usage found in mega-corpora are offered. Corpus linguistics has brought an entirely new level of research to lexical selection debates. Thanks to largely increased computer capabilities and databases, empirical data, in addition to expert opinion can be relied on when inquiry about lexical selection is made. This methodology, of course, has increased the validity of vocabulary research findings. Meyer (2002) commented on the creation of corpora.

> If corpus linguists understand the methodological assumptions underlying both the creation and subsequent analysis of a corpus, not only will they be able to create better corpora but they will be better able to judge whether the corpora they choose to analyze are valid for the particular linguistic analysis they wish to conduct. (p. xiv)

This study aimed at specifying and counting the general language that will prepare learners to understand and communicate at an advanced level in a specialized field, and was informed by frequency-based corpus data.

Summary

French vocabulary profile authors have stated that the profiles are meant to be suggestions for minimal competency at each level. They represent minimal cumulative lists of set notions teachers, textbooks' designers, evaluators, publishers have drawn upon to teach and evaluate learners. Unlike the English vocabulary profiles, they did not result from tested studies but rather didactic expertise. Though not prescriptive in nature, they tend to become a norm referred to for language achievement and proficiency.

Even if authors of the French vocabulary profiles suggested that vocabulary selection presented in the profiles should only be a central source of reference and not a standard, profile content and framework are being used for institutional teaching and learning, and assessment, evaluation. They are means of establishing equivalences with other proficiency systems around the globe through linking efforts of language exams to the CEFR "Can-Do" scale for instance. Learners' self-assessment, via instruments such as the Language Portfolio, leading to the creation of individually-based language proficiency norm over time is insufficient to measure vocabulary growth. Vocabulary ranges differ by level and by language. Vocabulary growth does not correspond with vocabulary range by level, and counting of words is problematic.

There were three reasons for trying to compare language levels using vocabulary size and vocabulary knowledge as indicators. One was that the original work on the framework included wordlists which gave some idea as to what vocabulary knowledge was expected of learners at some of the levels. A second was that vocabulary size and coverage are strongly connected to

vocabulary growth and ought to be a good general indicator for general word knowledge and language performance. Research bears out this idea (for example, Milton, 2006) and vocabulary size does, indeed, appear a good general indicator of foreign language ability. The third reason was that vocabulary is countable. It is possible to attach a figure to levels of knowledge and achievement. This fact allows comparisons between languages to be made, so it is possible to compare knowledge between English and French as foreign languages, for example, in a way which was not usually possible (Milton, 2006). Existing comparisons for French vocabulary acquisition in Britain provided convincing data of progressive decline in the knowledge of learners and the standard of school examinations over a period of decades (Milton, 2008).

This study followed in the footsteps of Milton's work. He compared vocabulary growth estimates of British learners of French to Baudot's 1992 frequency dictionary of French based on a corpus of around one million words. Herein the actual vocabulary content and size of the CEFR French vocabulary profiles offered by Beacco and colleagues were compared to the frequency data of close to a quarter of a million frequency dictionary of French corpus, and a close to one billion French newswire corpus.

This review showed that subjective task-based communicative vocabulary selections influenced by unclear theoretical underpinnings and variable language use should be complemented by more objective usage-based vocabulary selections. The result would be to better define the vocabulary range learners and teachers need to know to prepare themselves for advanced studies and professional work in French.

Chapter 3: Method

The purpose of this study was to analyze and substantiate the content of the CEFR French profiles as a resource for selecting vocabulary at various levels of French language proficiency. To do this the CEFR profiles were compared to the FDF and the FGC in order to determine the extent to which these three primary resources of commonly used French words overlap. The general methods used to accomplish this task are discussed in this chapter including research design, data sources, collection and organization of data, and data analysis.

Research Approach

The research design used for this study was primarily descriptive. The study was designed to identify the overlap (or commonality) between the profile lists of generally used French vocabulary. In addition, a negative case analysis was conducted to analyze those words that were not found to be in common between these resources. In order to accomplish these tasks, the study described the commonality of content between these lists using frequency analysis, the correlation of word frequency rankings, and a qualitative categorization analysis of words common to all three sources, common to only two of the three, or unique to a specific resource.

Salient Features of the Resources

The three primary French language resources used in this study were the CEFR, FDF, and the FGC (see their description in Chapter 2). Table 17 lists salient features that characterize each resource. All three resources cover written and spoken language. Their sizes ranged from 6,000 to 7,000 words to almost 1 billion. They were each organized in a unique and different way. The CEFR profiles are minimal word selections based on a notions/tasks framework and organized by proficiency levels. The FDF is a rank-ordered list of most frequent French lemmas

derived from a large corpus representing spoken and written French equally. The FGC is 15-year's worth of newswire text produced by two French press agencies.

Table 17

Salient Features of CEFR, FDF, and FGC Data

Features	CEFR	FDF	FGC
Register	spoken and written	spoken and written	spoken and written
Organization principles	general and specific notions; communicative tasks	half oral / half written texts	unstructured newswires
Size (words)	6,000 to 7,000	23 million	close to 1 billion
Corpus type	N/A	large	mega

This study's primary resources had mixed characteristics. However, each of the three resources were converted into comparable units of analysis. Table 18 highlights this point. Corpus-extracted words were converted into types. They were sorted and ranked by frequency. For FGC data, the threshold of 10 counts was adopted as the minimal criterion for inclusion in the frequency list from which FGC rankings were derived.

It was hoped that FDF lemmas would mostly, if not completely, overlap with CEFR lemmas since FDF lemmas come from a more representative corpus of most frequent French words. It was expected that less frequent FGC types would not overlap with CEFR and FDF types.

Data Preparation for Analysis

To get at actual profile and corpora lexical content, the vocabulary contained in French profiles level A1, A2, B1, B2, the FDF semi-large corpus, and the FGC mega corpus were converted into an analyzable and comparable electronic form. The analyses required the process of tokenization. This process allowed the identification of types.

Table 18

Data Characteristics of Primary Resources

	French Profiles (CEFR)	FDF	FGC (3rd edition)
Publication Dates	2004-2011	2009	2011
Lexical Categories	lemmas, types, multiword units	lemmas	newswire text
Conversion needed	types	types	types
Categorization by notions	yes	no	no

The vocabulary was analyzed as types according to the research approach. The presence of a word unit in a specific lexical source was coded dichotomously, either as being present with a "1" or absent with a "0". As Table 19 illustrates, primary data types were categorized into the following lexical units.

Table 19

Primary Resources Subdivisions and Codes

Resources	Subdivisions and Codes			
CEFR	A1 types with code 1 or 0	A2 types with code 1 or 0	B1 types with code 1 or 0	B2 types with code 1 or 0
FDF	FDF types with code 1 or 0			
FGC	FGC types with code 1 or 0			

Converting CEFR profiles into types. Multiple word units present in the CEFR were converted into *types*. A multi-word unit is defined as a vocabulary item which consists of a sequence of two or more word forms (a word being simply an orthographic unit). Each type extracted from each of the four CEFR French vocabulary profiles under scrutiny: *Le Niveau A*1, *Le Niveau A2*, *Le Niveau B1*, and *Le Niveau B2*, was entered into a master vocabulary Access database from each profile's word index found at the end of each book. They were compared to

the list found under general and specific notions chapters of profile publications for verification purposes.

For greater ease and speed, Access-formatted data were imported into an Excel spreadsheet. A set of Perl and command-line scripts to sort, compare, analyze, rank, and list the results of this study was developed. The BDLex program for part of speech (POS) tagging was used to convert lemmas into types. Additional data cleaning and spot checking was done by hand. Each CEFR type was linked to its corresponding FDF ranking and/or FGC ranking, if available.

Eliminating type overlap internal to CEFR profiles. In order to assist the comparison of types, an initial analysis of the CEFR profiles was conducted to identify vocabulary first introduced at level A1, A2, B1, and B2. This procedure allowed comparison with the FDF and the FGC. It was also performed to ascertain that no overlap existed internal to the CEFR profiles and to limit the final vocabulary count to unique wordforms in each resource. Some words were multi-word units such as *aller-retour* (round trip), *boîte à lettres* (mailbox), *il y a* (there is), *permis de conduire* (driver license), *avoir des enfants* (to have children). Some single words were inflected types such as *toilettes* (restroom), *échecs* (chess), *olympiques* (olympic), *petite* in petite-fille (grand-daughter), *rendez* in rendez-vous (appointment). Other single words were lemmas such as *valider* (validate), *savourer* (savor), *travail* (work), *rugueux* (rough), *plus* (more). Once multi-word expressions were reduced to single word units, a comparison was done across and within levels to check whether the same wordforms were introduced more than once. This comparison led to the discovery that wordform overlap existed within and across CEFR proficiency levels. Table 20 illustrates the problem of overlap internal to the profiles. Wordforms were presented at four, three, two, or only one of the four levels. When a wordform was

Table 20

Uniqueness and Overlap Found in the CEFR Profiles

Wordforms	Count	Examples
Unique to level A1	26	*bonjour, bonsoir, bravo, comment, entendu, milliard, millier, non, oui, pardon, plaît, politique, revoir, salut, zéro*
Unique to level A2	21	*fêtes, photographie, avancer, débutant, séparé*
Unique to level B1	243	*inflammatoire, appellations, arabophone, artichaut*
Unique to level B2	3,035	*boulot, empêcher, longer, lotion, macédoine*
Present in level A1, A2, B1, & B2	888	*rond, ronde, rose, rouge, rousse, route, roux, rue*
Present in level A1 & A2	930	*accent, accident, accord, accueil, acheter, addition, adjectif, adresse, adulte*
Present in level A1, A2, & B1	927	*bagage, baguette, bain, banane, bancaire, banlieue, banque, bas*
Present in level A1 & B1	952	*chaîne, chaise, chambre, champ, change, changer, chanter, chapeau*
Present in level A1 & B2	910	*déjeuner, demain, demander, demi, démocratie, dent, dentifrice, dentiste, dents*
Present in level A2 & B1	1,590	*écouter, écrire, écrit, écrite, égal, égale, égalité, église, élection, électricité*
Present in level A2, B1, & B2	1,476	*février, fiche, fièvre, fille, film, fils, fin, fleur, fleuve, fois, foncé*
Present in level A2 & B2	1,483	*gant, garage, garçon, garder, gare, gâteau, gâteaux, gauche, gaz*
Present in level B1 & B2	3,012	*heure, heureuse, heureux, hier, hifi, hindouiste, histoire, hiver, homme, honnête,*

presented more than once, the first occurrence of the wordform at a given proficiency level was the one that was used and counted. The other listings of that same wordform were not counted. For instance, out of the entire number of wordforms at level A1, only 26 were unique to that level.

Converting the Frequency Dictionary of French into types. The production of the FDF corpus and target lemmatized frequency vocabulary already entailed several steps which have been described in the introduction section of the FDF: (a) text selection, (b) corpus standardization and annotation, (c) target vocabulary identification and description, and (d) development of associated information which include a combination of frequency and dispersion information. The text collection of the FDF corpus involved work in corpus standardization or pre-processing.

The same set of Perl and command-line scripts to sort, compare, analyze, rank, and list results was used for FDF results. There also, the BDLex program was used to convert lemmas into types. Using BDLex to convert lemmas into types meant that frequency and rank order information was only available for lemmas and not for their associated inflections.

Converting the French Gigaword Corpus into types. The extremely large FGC is constituted of streams of documents with SGML (Standard Generalized Markup Language) parts of speech (POS) tags. Tree-Tagger was the tool used for annotating text with POS and lemma information. This tool has also been used to tag languages such as German, Spanish, Russian, Greek, Chinese, Swahili, or Estonian texts. Tags serve to mark document boundaries (e.g. individual news stories), as well as important components or divisions within each document (unique document identifiers, headlines, paragraph boundaries). Markup formatting was kept simple (shallow) to facilitate the automatic SGML tagging process, and/or filter the data to retain

or discard text content according to particular research needs. Perl script was used to remove SGML tags. The FGC data underwent consistent quality control to eliminate various forms of errors. Even though the FGC data was corrected, some errors remained such as spelling or tagging errors like miscounting hyphenated words as one word instead of two. Given the considerable size of the resource, no attempt was made to address this property of the data.

The same Perl, command-line scripts, and BDLex program were used for FGC to accomplish the same end. Additional data cleaning and spot checking was done by hand. Frequency and ranking information was obtained. However, a dispersion coefficient was not calculated since Gigaword only represents the newswire register, and thus would likely not yield important information.

Determining overall quantified compilation. The total possible number of types was obtained for the CEFR, FDF, and FGC and used to calculate percentages for each section delimited by this study's research question. In the study's three primary resources, 369,607 unique types were identified., Out of these, 11.2% were found in the CEFR. Of those CEFR types, 14.2% were level A1 types, 12.7% level A2, 28.9% level B1, and 44.2% level B2 types respectively. After the preliminary CEFR profile analysis shown in Table 20, Table 21 summarizes unique single wordform totals for each proficiency level. It also lists the actual number of types per proficiency level derived from these wordforms.

The FDF 5,000 most frequent lemmas represent close to 18.1 million tokens (78.6%) of the frequency dictionary 23 million token corpus. Once inflected with the BDLex program, the 5,000 lemmas yielded 53,511 types, or an average of 11 inflections per French lemma. A total of 337,661 types in the FGC are considered in this study from an initial total of 2,608,706 types.

This result was obtained after handcleaning for strings that contained only numbers, or strange formats or characters, and the use of the 10 count threshold for inclusion (see Table 22).

Table 21

Number of CEFR Initial One-Word Units, Unique Wordforms, and Derived Types by Proficiency Level at Which They Were First Introduced

Proficiency Level	Initial One-Word Unit Count	Unique Wordforms First Introduced at	Types Derived from Unique Wordforms
A1	975	975	5,853
A2	1,599	670	5,234
B1	3,354	1,743	11,913
B2	6,034	3,019	18,255
Total	11,962	6,407	41,255

Table 22

Total Number of Types by Primary Resource

Types	Type Count
CEFR Profiles	41,255
FDF	53,511
FGC*	337,661
Total Count	369,607

* threshold of 10 counts for inclusion

Planned Analyses

The idenfication of total number of types by resource allowed the calculation of unique types per study section. It also permitted to ascertain how frequent these types were. Frequency data was not available for the CEFR vocabulary profiles. However, frequencies and frequency rankings were already available for the 5,000 most frequent FDF lemmas. Frequencies and frequency rankings were determined for FGC types. The total possible number of unique types

for the CEFR, FDF, and FGC was established. This determination made possible the computation of types per study section and respective percentages. The seven data sections illustrated in Figure 1 (page 7) were defined and analyzed as follows:

1. The core of types common to the CEFR, FDF, and FGC;

then, the three sections of partial overlap including

2. Types common only to the FDF and FGC;

3. Types common only to the CEFR and FDF, and

4. Types common only to the CEFR and FGC;

and finally, the three sections of no overlap including

5. Types unique to the FDF,

6. Types unique to the FGC, and

7. Types unique to the CEFR.

Table 23 and Figure 3 show that, of all the 369,607 unique types considered in this study, six of the seven study sections cover only 18% of all types; the remainder is part of types unique to the FGC.

For each section mentioned above, French types were quantified and qualitatively described. They were grouped by proficiency levels (A1, A2, B1, B2). Existing FDF and FGC frequency rankings were tied to them. Using available paired FDF and FGC frequency rankings, Spearman rho correlation coefficients were also calculated to summarize the direction and strength of association of types endowed with such rankings in the study's common core and the partial FDF/FGC overlap. In addition, the discrepancy between FDF/FGC frequency rankings was computed. Types were ordered on a least to most discrepant scale. Least and most discrepant types were described within proficiency levels, if available. These analyses gave actual quantities

of words by proficiency levels. They allowed a commentary on vocabulary quantity, quality, and growth rate at lower proficiency levels.

Table 23

Number of Types in Primary Resources by Study Section

Section	Description	Unique Types	Percentage
1	Core common to CEFR, FDF, and FGC	16,649	4.5
2	Common only to FDF and FGC	11,649	3.1
3	Common only to CEFR and FDF	12,056	3.3
4	Common only to CEFR and FGC	5,817	1.6
5	Unique to FDF	13,157	3.6
6	Unique to FGC	303,546	82.1
7	Unique to CEFR	6,733	1.8
Total		369,607	100.0

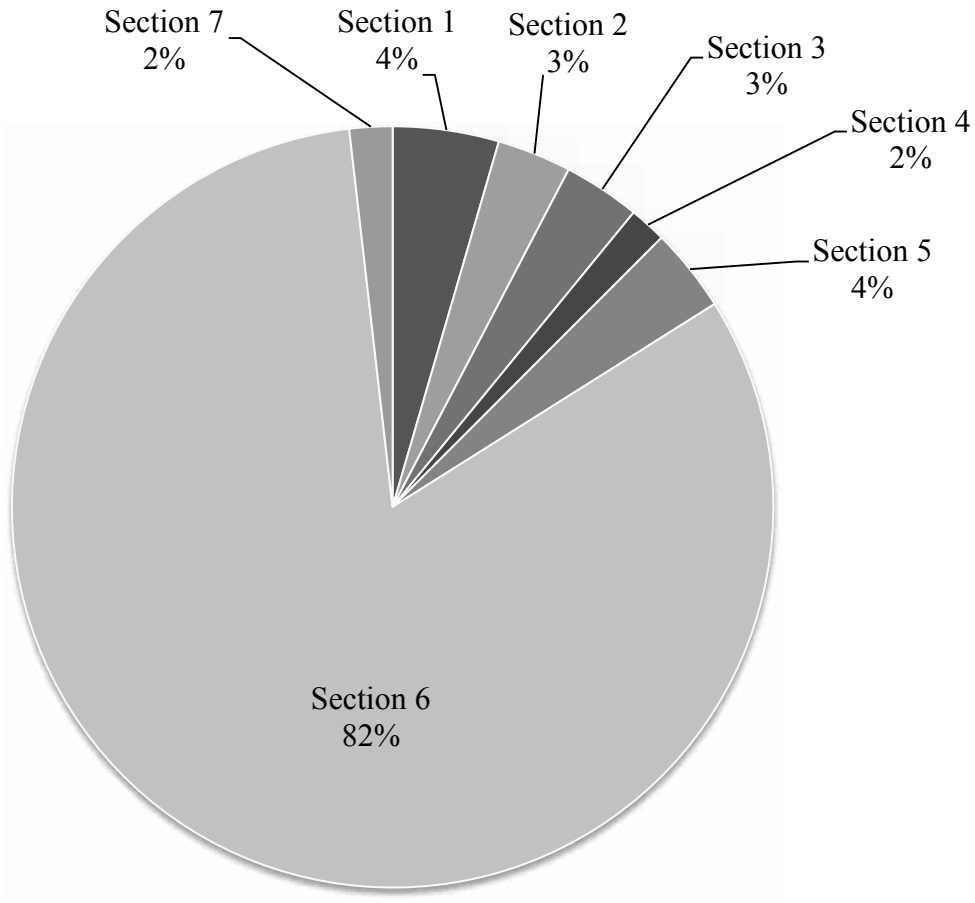

Figure 3. Types by study sections

Chapter 4: Results and Recommendations

Each of the seven sections outlined in Chapter 3 were analyzed and described quantitatively and qualitatively. The findings deal with lexical types unique to each of these seven sections. Modifications to the initial CEFR profile selections are based on usage. They are recommended by proficiency levels, when available, and are synthesized at the end of the chapter.

Overview of Findings by Resources

The initial analysis was done to get the big picture of how the CEFR profiles compare with the FDF and FGC types included in this study. Table 24 breaks down CEFR types by proficiency level and overlap areas. Two out of five (40.4%) CEFR types belonged to the common core. Overlap with FDF was 29.2% and overlap with FGC was 14.1%, and 16.3% were unique to the CEFR profiles.

Table 24

CEFR Types by Proficiency Level and Degree of Overlap with Other Resources

Proficiency Level	Unique to CEFR	Common to CEFR, FDF & FGC	Common Only to CEFR & FDF	Common Only to CEFR & FGC	Total
A1	126	3,859	1,613	255	5,853
A2	354	2,658	1,816	406	5,234
B1	1,585	5,214	3,865	1,249	11,913
B2	4,668	4,918	4,762	3,907	18,255
Total	6,733	16,649	12,056	5,817	41,255
Percent	16.3	40.4	29.2	14.1	100.00

Results were organized according to inclusion priorities determined by type commonality, frequency, and frequency rankings obtained from the FDF and FGC corpora. No such information was provided for CEFR types. Priorities for inclusion and exclusion are shown

on Table 25. One (1) means that the section in question should be considered first and seven (7) that it should be considered last for inclusion or exclusion whatever the case may be.

Table 25

Priorities for Inclusion and Exclusion in CEFR Profiles

Order of Inclusion	Order of Exclusion	Type Sources
1	7	Common to CEFR, FDF, and FGC (core)
2	6	Common only to FDF and FGC
3	5	Common only to CEFR and FDF
4	4	Common only to CEFR and FGC
5	3	Unique to FDF
6	2	Unique to FGC
7	1	Unique to CEFR

Judgment regarding acceptable/questionable types was also based on the presence of *passé simple* or exclusively subjunctive forms in the resources' subgroups. Their presence was partly an artifact of generating all possible inflected types of CEFR and FDF lexical units. Numerous *passé simple* and exclusively subjunctive forms are infrequent and used at advanced linguistic level. When they occurred in non-core sections with no frequency ranking, these forms were given a negative qualitative assessment. Other verbal forms received a positive qualitative assessment. No stricter norm was used for qualitative assessment since the study was exploratory and limited to forms, not meanings. Criteria specifications will be noted in the chapter when used.

Spearman rho correlations were computed to support inclusion/non-inclusion recommendations for overhauled CEFR profiles. They confirmed that a positive association existed between all types endowed with paired FDF and FGC frequency rankings, those part of

the common core, or common only to FDF and FGC. Paired rankings were available for 4,902 types that produced a Spearman rho correlation coefficient of .781, significant at the 0.01 level (2-tailed). These types were subdivided into two groups: 2,904 common core types (Spearman rho = .819), and 1,998 types common only to the FDF and FGC (Spearman rho = .639). They were sorted in increasing order of ranking discrepancy. The absolute value of their rank difference ranged from zero (least discrepant) to 129,599 (most discrepant). Types with discrepancy spanning from 0 to 100 (n = 341) correlated at .989, a strong positive association significant at the 0.01 level (2-tailed). Types whose discrepancy varied from 10,002 to 129,599 (n = 198) correlated at .142, a weak positive association significant at the 0.05 level (2-tailed). The smaller the absolute value of the ranking difference, the stronger the positive rank association; and, the larger this value, the weaker their positive association. Findings are detailed below in relevant sections on the common core, and on types common only to the FDF and the FGC where least and most discrepant groups are described. As a general rule, types with strong positive association should be learned first, those with weak positive association last. However, they are, as a whole, more frequent than the other types of the three resources considered.

Types Found in Resources by Study Sections

Findings and recommendations are now presented according to the seven sections outlined previously. They follow the determined inclusion priority sequence, i.e. first, CEFR types common to all three resources, then very frequent FDF and FGC types not overlapping with CEFR types, followed by CEFR types overlapping with one or the other of the two contemporary corpus resources, and finally FDF, FGC, and CEFR types unique to their respective resources.

Common to the CEFR, FDF, and FGC. The common core (depicted in Figure 4) included 16,649 types (4.5%) of the total 369,607 unique types identified in the CEFR, FDF, and FGC. Core types are the first pool to draw from for inclusion in the overhauled CEFR profiles. They should be given first learning and teaching priority since they appear in all three resources.

Core types represented common actions such as

aimer, comprendre, marcher, faire, penser, travailler.

They also described people, animals, and things such as

enfant, homme, femme, animal, chat, poule, choses, fleur, bicyclette.

They contained an average of 23 verbal inflections per infinitive lemma which comes as no surprise since this number of inflections can go up to 65 for French verbs.

A breakdown of core types by proficiency level is presented in Table 26. Differences between the total number of CEFR types and core types existed at each proficiency level. It is important to recognize that three out of five CEFR types (24,606 types, 59.6%) were not present in the common core. Initially, CEFR A1 non-core types represented 34.1%, then the percentage jumped to 49.2% for A2 non-core types, to 56.3% for B1, and it reached 73.1% for CEFR B2 non-core types. The greater the percentage difference, the greater the likelihood for CEFR types to be infrequent. The highest number of CEFR non-core types was offered at intermediate proficiency when over three times as many types were introduced for level B2 as for level A1.

Table 26
Number and Percentage of Common Core Types by CEFR Proficiency Levels

Proficiency Level	Total Number of CEFR Types	Common to CEFR, FDF, and FGC	
		Number	Percentage
A1	5,853	3,859	65.9
A2	5,234	2,658	50.8
B1	11,913	5,214	43.7
B2	18,255	4,918	26.9
Total	41,255	16,649	40.4

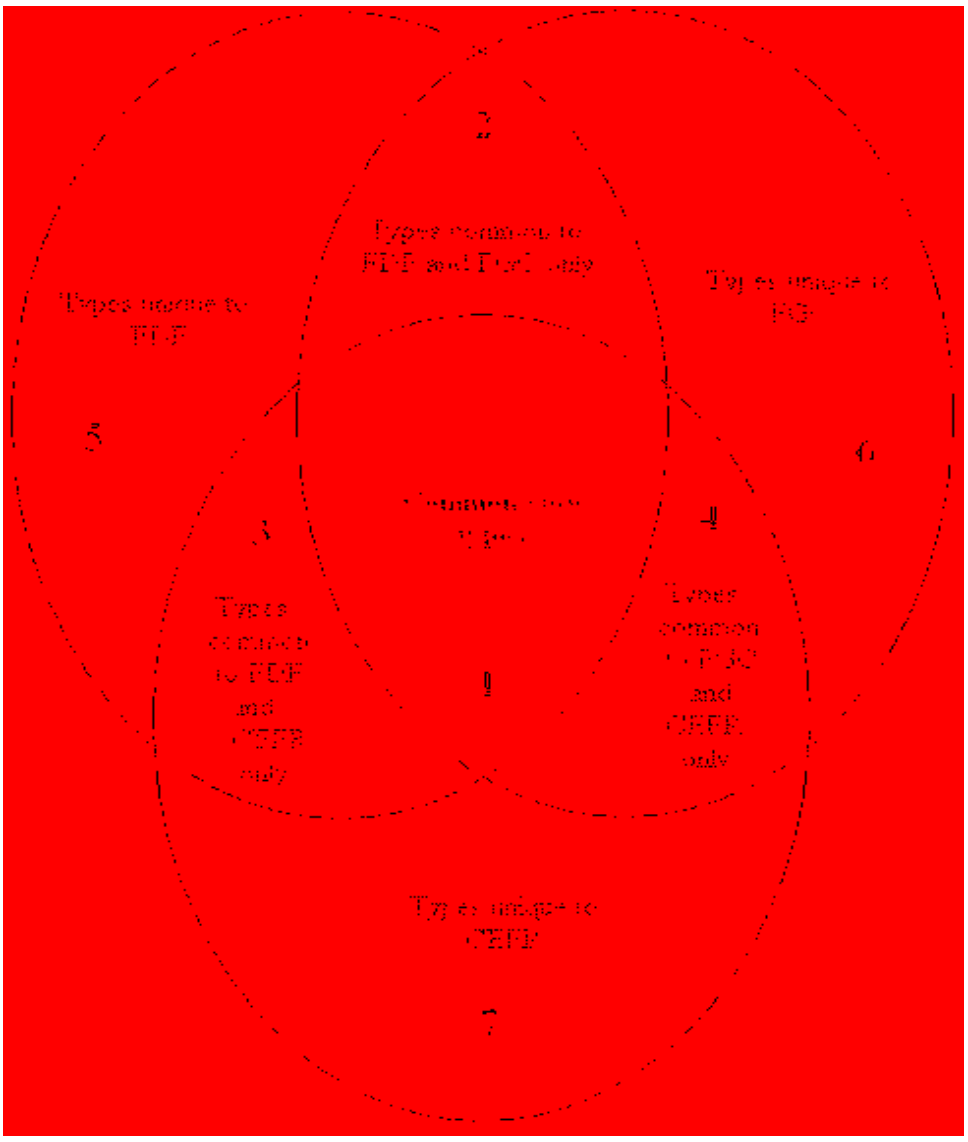

Figure 4. Types common to the CEFR, FDF, and FGC

Core types, although all acceptable since they belonged to the most frequent words of the French language, could still be categorized as acceptable and less acceptable. Criteria of acceptability at proficiency level A1 to B2 were determined. FDF ranking based on frequency and dispersion, FGC ranking based on frequency, and the absolute value of the difference between FDF and FGC rankings allowed for degrees of acceptability by proficiency level. In addition, the Spearman rho correlation coefficients for the 789 core paired rankings at level A1 reached .850; for the 407 core A2 paired rankings, .796; for the 937 core B1 paired rankings, .799; and for the 771 core B2 paired rankings, .733, all displaying a strong positive association. On one hand, 270 out of 341 (79.4%) least discrepant types were part of the common core. At level A1 were 147, 32 at A2, 53 at B1, and 38 at B2, respectively. On the other, 66 (33% of 198 total) most discrepant types also belonged to the common core, i.e. 11 at level A1, 3 at A2, 22 at B1, and 30 at B2.

Table 27 accounts for core types by inclusion criteria in decreasing order of acceptability, and by proficiency level. Core types not only constituted 40.4% of CEFR types, the findings further showed that 78.3% of them were inflections of the 5,000 most frequent French lemmas. These inflections were, however, less frequently used in the FGC (see Table 27 criterion "**d**").

Acceptable core A1 types. All the 3,859 core A1 types were acceptable for inclusion in the CEFR profiles. Nevertheless, learners would benefit from early instruction of some over others based on frequency and ranking criteria. Criterion "**a**" core types should be taught first, criterion "**d**" core types last, or likely moved to the A2 proficiency level.

Table 27

Acceptable Core Types by Inclusion Criteria

Inclusion Criteria	Description	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
a	Δ|FDF rank -FGC rank| ≤ 5,000	766	389	854	672	2,681
b	FDF ranks ≤ 5,000 or FGC ranks ≤ 5,000	39	26	102	121	288
c	FGC ranks > 5,000	147	114	196	191	648
d	FGC ranks > 216,059	2,907	2,129	4,062	3,934	13,032
	Total	3,859	2,658	5,214	4,918	16,649

Core types first introduced at level A1 with paired FDF and FGC rankings whose difference did not exceed the absolute value of 5,000 should be kept and taught at level A1 consistent with the increasing difference value (see Table 27 criterion "**a**"). Core A1 types meeting criterion "**a**" were most frequent function words and infinitives of the French language such as

le, un, de, en, y, très, être, avoir, pouvoir, mettre, dire, devoir.

Also acceptable were core A1 types with an FDF or an FGC ranking of 1 to 5,000 (see Table 27 criterion "**b**"). These types were greeting or congratulatory forms, feminine, masculine, and plural forms related to both general and specific CEFR semantic notions, e.g.

bonsoir, mademoiselle, bravo, *première, correspondance, gentil, copain, êtres, parents*.

Another instance of acceptable A1 core types would be core types introduced at other proficiency levels, i.e., A2, B1, or B2, when the A1 type threshold needed for instruction was not reached.

Core A1 types with FGC ranking above 5,000 should be considered less acceptable, meaning they should be taught or learned last (see Table 27 criterion "**c**"). Less acceptable core A1 types were words such as names of female domestic animals which could be taken literally or figuratively such as

chienne, chatte, cochonne,

or irregular forms with spelling differences between *bel* and *belle*, or forms ending with the suffix *–ant* (similar to the English suffix *–ing*) found in types such as

appelant, apprenant, croyant

which are not used as in the English language to form the present tense, and can refer to nominal, adjectival, and/or verbal forms.

Another category of less acceptable core A1 types came from the group with no FGC ranking or ranking beyond the 216,059th FGC rank (see Table 27 criterion "**d**"). Core A1 types of this category were, for instance, forms that morphologically look like French plurals such as

ans, avants, avoirs, biens, eaux, étés, travaux, yeux,

but could sometimes be nominal or verbal forms such as

écrits, faits, lions,

infrequent feminine forms such as

bleue and *feue*,

or some less and less frequent subjunctive forms, e.g.

aie, aille, mette,

and almost all the inflections of infinitives like

finir or *asseoir.*

Description of discrepant core A1 types. Least discrepant A1 types (discrepancy value = 0 to 99) were function words such as prepositional, determinant, conjunctive forms, e.g.

de, le, par,avec, dans, que, en, entre, et, mais, pour, un, vers, pas, plus, sur, alors, sous, ou, sans, avant, comme, très, moins, même, depuis, après, peu, devant, encore, près, bien, toujours, soit;

pronominal forms i.e.

se, il, y;

adjectival forms i.e.

second, dix, jeune, nouveau, tout, rouge, deux, premier, grand, important, dernier, trois, social, bleu, petit, vert, mise;

and verbal infinitives i.e.

faire, avoir, devoir, être, pouvoir, prononcer, aller, mourir, prendre, apprendre, rester, dire, mettre, naître, arrêter, passer, sortir, recevoir, appeler, venir, demander, partir, voir, décider.

The least discrepant core nominal forms that surfaced were

retour, départ, route, usine, monsieur, point, place, son, cours, heure, banque, fin, démocratie, politique, face, année, député, fille, mesure, monde, étudiant, identité, femme, jour, droit, personne, couple, hauteur, nombre, bébé, vent, autobus, médicament, corps, enfant, eau, loi, entreprise, an, pays, maison, homme, information, état, service, fils, mot, étage, but, rue, frère, situation, numéro, partie, travail, restaurant, jeu, famille, lieu, mois, temps, voisin, fois, photo, tente, côté, gouvernement

listed in decreasing order of discrepancy. Most discrepant A1 core types in decreasing order were

tiens, bonsoir, oh, bravo, veuf, dictionnaire, arrivant, mademoiselle, dedans, bonjour, coton.

As seen above, core A1 types were very common. These most frequent forms used as function words and for every day actions were found in both general and specific semantic CEFR notions. This finding is not surprising given the frequency of use of these words.

Acceptable core A2 types. All the 2,658 core A2 types were acceptable for inclusion in the CEFR profiles. Similarly to core A1 types, some should, however, be presented to the learner earlier than others. This decision should be based on frequency and ranking criteria allowing the

identification of acceptable core A2 types that should stay at this level. Less acceptable A2 types should be taught last or be moved to the B1 proficiency level.

Core A2 types with FDF and FGC ranking whose difference did not exceed the absolute value of 5,000 should be kept and taught at level A2. This criterion is consistent with the increasing difference value, and the maximum number of types set for learning and teaching at level A1 (see Table 27 criterion "**a**"). Core A2 types meeting criterion "**a**" were again among the most frequent French types, most common forms representing actions and the way they are performed such as

trouver, porter, aider, rapidement, doucement;

forms representing functions occupied in society such as

commerçant, directeur, maire,

forms alluding to geography and dwelling places such as

planète, continent, province;

forms alluding to life and feelings such as

santé, vie, maternité, peur, amoureux, malheureux;

or forms referring to nationalities and languages such as

français, anglais, allemand.

Illustrative core A2 types with a FDF or FGC ranking of 1 to 5,000 (see Table 27 criterion "**b**") were forms that, even though frequent, did not occur as frequently in the FGC such as

librairie, génial, gai, connecter, quelquefois,

possibly because people buy less and less books in a *librairie* (*bookstore*), they associate *gai* to a lifestyle, and use the English spelling *gay;* other forms were common masculine forms such as

correct, mou, amusant, élégant

or feminine forms such as

lèvre, correction, tante, goutte, forces, règles,

or infinitive forms

coller, taper, grossir, déménager, copier.

Core A2 types whose FGC ranking was higher than 5,000, should be less acceptable (see Table 27 criterion "**c**"). They are feminine forms of occupations and nationalities, perhaps due to predominant French use of the masculine in case both genders are represented, such as

maîtresse, patronne, doctoresse, anglaise, polonaise.

Also, as for core A1 types, forms ending with the suffix *–ant* were noticeable such as

cuisant, aidant, grossissant.

In addition, at level A2, a good number of masculine forms ending in *–é* appeared such as

composé, préféré, habillé;

and corresponding feminine forms

jetée, cuite, couverte;

forms most often used to conjugate the third person singular of the present tense such as

c*asse, pousse, traverse, rit*.

Further, core A2 types with FGC rankings higher than 216,059 should also belong to the less acceptable A2 category (see Table 27 criterion "**d**"). Of the 103 common French verbs and 382 other forms (part of the core A2 type pool) were 1,751 inflections falling into the criterion "**d**" category. Illustrative examples follow. Among verbal inflections, the average was 18 forms per verbal lemma such as

accepta, acceptaient, acceptais, acceptait, accepte, accepté, acceptée, acceptées, acceptent, acceptera, accepterai, accepteraient, accepterais, accepterait, acceptèrent, accepterez, accepterions, accepterons, accepteront, acceptes, acceptés, acceptez, acceptiez, acceptions, acceptons.

Examples of these common lemmas included

adorer, calculer, détester, enseigner, guérir, marier, oublier, présenter.

They contained mainly feminine and plural terms describing human associations, occupations, and nationalities such as

cousine, chanteuses, chômeurs, espagnole;

singular feminine forms such as

gaie, molle, nulle;

plural nominal forms such as

achats, balles, émissions, factures, guerres, histoires, jouets, lèvres, matières;

and other plural adjectival forms such as

agréables, claires, faciles, olympiques, spéciaux.

Description of discrepant core A2 types. Least discrepant A2 core types, in increasing order of discrepancy, were as follows (discrepancy value = 2 to 98), i.e.

projet, retrouver, présenter, aide, professionnel, contrat, coup, autre, renseignement, chiffre, employé, région, où, public, rendre, santé, si, aussi, installer, guerre, enlever, trouver, rencontrer, religieux, musée, plage, inscrire, vie, artiste, vente, vouloir, donner.

Most discrepant A2 core types in decreasing order are, i.e.

quelquefois, amusant, moyenne;

Core A2 types seemed to be more representative of some specific notions. They related to trades and occupations, geography, feelings, social life, human characteristics, and nationality. An increasing number of feminine, infinitive, participial, third-person singular present, plural forms were witnessed. In addition, a great number of verbal inflections with some more commonly used passé simple forms were counted. These findings might help decide where to first focus attention in the learning and teaching process.

Acceptable core B1 types. All the 5,214 core B1 types were also acceptable for inclusion in the CEFR profiles. As previously mentioned, some should however be presented to the learner earlier than others based on frequency and ranking criteria which determine the sequence in which they should be taught, first, last, or moved to a higher proficiency level.

Core B1 types with paired FDF and FGC rankings whose difference did not exceed the absolute value of 5,000 should be kept and taught at level B1 consistent with the increasing difference value (see Table 27 criterion "**a**"). Examples of core level B1 types meeting criterion "**a**" were still among most frequent French types such as words qualifying occupations and their associated inflections, e.g.

écrivain, conservateur, fonctionnaire;

words qualifying the religious and ethical realm, and some of their associated inflections, i.e.,

juif, temple, prière, mensonge, respecter, esprit, sacré;

words using the suffix –*ion* indicating a process and associated inflected forms, e.g.,

augmentation, augmenter, multiplication, multiplier, formation, former;

types referring to technology and means of communication, e.g.

satellite, médias, slogan;

types usually belonging to the political or judicial register, e.g.

débat, parlement, référendum;

additional types describing culture and language, e.g.

japonais, francophone, langage;

types describing qualities such as

prudence, douceur, générosité, envie;

or describing anatomical parts, e.g.

cou, os, organe;

types denoting positive values, e.g.

progrès, enthousiasme, espoir;

or negative values, e.g.

violence, stress, mécontentement;

number and order types:

cinq, cinquième, secondaire;

types indicating how things are performed with the suffix *–ment* (equivalent of the English suffix *–ly*) and based on adjectival forms, e.g.

habituellement, rarement, exactement.

Other acceptable core B1 types were those with a FDF or FGC ranking of 1 to 5,000 (see Table 27 criterion "**b**"). Illustrative examples are infinitive forms, e.g.

décharger, questionner, exagérer;

epicene adjectival forms such as

horrible, énergique, souple;

nominal forms ending with *–té* such as

rapidité, obscurité, simplicité;

feminine nominal forms such as

honte, vieillesse, indifférence

denoting qualities and feelings; eclectic masculine nominal forms such as

panier, noyau, angle, désespoir, socialisme, blé, emplacement, raisonnement, bouton, interprète;

forms ending with *–ant* such as

exigeant, passionnant, composant;

the masculine adjectival forms such as

banal, sourd, pair, rationnel;

and types which, although frequent, seem to be more technical or less common e.g.

poil, occurence, réformiste.

A class of less acceptable core B1 types were those whose FGC ranking is above 5,000. (see Table 27 criterion "**c**"). Examples of this category of core B1 types are the feminine types such as

hôtesse, jumelle, prêtresse, demanderesse, raie, louve

which represent less common inflections of types already introduced at lower proficiency levels; feminine adjectival forms

régulière, nette, générale, manuelle;

and a few plural forms

amers, actifs, minima;

types mainly used for the conjugation of the third person singular of the present tense but also often used as nouns ending with *–e*,

dispute, décharge, relève, conserve;

and a few emerging *passé simple*-like forms, i.e.

versa, serra, fias.

Many past participle forms were also found in this category, e.g.

promis, fondu, ressenti, nourri, enveloppé, amenée, composées;

or present participle forms with the *–ant* suffix and their inflections, e.g.

gérant, pratiquant, négociant, exposant, penchant, signifiant,

types which often have drifted or are drifting semantically.

Another class of less acceptable core B1 types were those whose FGC ranking lied beyond the 216,059th rank (see Table 27 criterion "**d**"). Core B1 types in this category were mainly an average of 17 inflections per verbal lemma (3,259 total) of 193 infinitives such as

appartenir, circuler, découvrir, envelopper, justifier, loger, multiplier, nommer, reconnaître, terminer;

the rest of the pool contained plural forms (762) such as

alimentations, affaires, automobiles, crédits, énormes, familiaux;

and singular forms (33) which were predominantly feminine, e.g.

acheteuse, contemporaine, employeuse, fatiguée, nerveuse, vitale.

Description of discrepant core B1 types. B1 least discrepant types were as follows, in increasing order of discrepancy (discrepancy value = 0 to 100), i.e.

conservateur, écrivain, réduction, afin, ce, durée, fonctionnaire, découvrir, qui, charge, ne, sol, promettre, commune, obtenir, catholique, charger, débat, négocier, poudre, mission, ouvrier, conseil, opposer, succès, quelque, satellite, édition, groupe, intention, médias, port, titre, conflit, discussion, huile, ailleurs, exprimer, allocation, cité, contre, fédéral, institut, scène, manifester, progresser, assurer, comité, durant, juger, plusieurs, carrière, revendication.

Most discrepant B1 core types in decreasing order were, i.e.

continuellement, vingtième, réfléchi, immobile, aimable, raisonnement, raide, typique, débrouiller, commencement, exposé, deviner, rationnel, vocabulaire, nuance, semblant, réformiste, philosophique, déduire, courbe, productif, ennuyer.

In addition to observations already mentioned about core A2 types, a greater variety of inflections, affixes, adjectival, and verbal inflections was seen at this level. Further, specific notions were better represented than general notions, e.g. occupations, religion, communication, politics, culture and language, feelings, and anatomy vs. quantities or logical relations.

Acceptable core B2 types. Again, all the 4,918 core B2 types were acceptable for inclusion in the CEFR profiles. Some should be presented to the learner earlier than others based on frequency and ranking criteria. Core B2 types with paired FDF and FGC rankings, whose difference did not exceed the absolute value of 5,000 should be kept and taught at level B2. They should also consistently follow the order of increasing difference value. Core B2 types meeting

criterion "**a**" (see Table 27) continued to be among the most frequent French types. They were used in diverse domains, among others, religion, e.g.

évêque, islam, culte,

occupations, hobbies, and sports, e.g.

militaire, chasseur, football,

health and anatomy, e.g.

symptôme, crâne,

work and the economy, e.g.

embauche, capital, épargne,

technology, e.g.

virtuel, écran,

music, e.g.

orchestre, rythme, instrument.

The acceptable core B2 forms also represented more abstract concepts such as

aveu, serment, censure, enjeu, homologue.

Among these 672 forms, close to one third (199) were infinitive forms, e.g.

réaliser, constituer, rétablir, émettre, prescrire, résulter, grimper.

Adverbial and prepositional forms were also present such as

ainsi, puis, toutefois, hors;

and more precisely also adverbial forms with the ending *–ment,* e.g.

particulièrement, éventuellement, génétiquement.

In addition, some forms ending with *–ant* were seen again*,* e.g.

stupéfiant, délinquant, militant;

feminine forms (58) ending with *–ion, -tion,* or *–sion*, denoting a process, e.g.

intervention, fusion, inflation, composition, confession, mutation;

adjectival forms e.g.

triple, notre, similaire, fidèle, franc, populaire;

feminine forms ending with *–ance, -ence,* or *–ense,* e.g.

performance, alliance, puissance, espérance, conséquence, défense, référence, récompense, influence, compétence, urgence;

masculine nominal forms ending with *–ment,* indicating a result, e.g.

ralentissement, document, remplacement, comportement, rapprochement, rétablissement, rassemblement, fondement;

feminine forms ending with *-té* e.g.

fermeté, autorité, culpabilité;

masculine forms ending with *-eau* e.g.

réseau, niveau, troupeau;

feminine forms ending with *–ie,* e.g.

énergie, hiérarchie, thérapie;

or masculine forms ending with *-if* ,e.g.

décisif, exécutif, négatif;

masculine, feminine, or epicene adjectival and/or nominal forms such as

saint, originaire, interne, mental, éternel, imminent, cruel;

feminine nominal forms ending with *–e*, e.g.

phase, expertise, cellule, corde;

or masculine nominal forms ending with *–e*, such as

semestre, disque, terme, intervalle.

Core B2 types whose FDF or FGC rank ranged between 1 to 5,000 should also be considered acceptable (see Table 27 criterion "**b**"). Some of these types became familiar or even vulgar forms such as

boulot, gueule, cul, ficher, crever;

others referred to ideas or disciplines such as

idéologie, rhétorique, linguistique;

or certain feelings such as

solitude, mépris, orgueil.

Forms less common such as

concept, ultérieur, subtil, faisceau

were also present. Other examples included infinitive forms such as

découler, tracer, baigner;

present and past participle forms e.g.

dominant, infini, débouché, inexistant, ressortissant;

masculine nominal forms e.g.

grandeur, rendement, support, dynamisme, voisinage;

feminine nominal forms such as

connexion, aptitude, équité;

adverbial forms such as

constamment, brusquement, fondamentalement;

and masculine or epicene adjectival forms such as

merveilleux, neutre, antique.

Core B2 types whose FGC rank was above the 5,000th rank should be considered less acceptable (see Table 27 criterion "**c**"). Illustrative examples of these less acceptable core B2 types are, in this case, present participle forms often used as nouns e.g.

remplaçant, exploitant, surveillant, dissolvant;

present participle forms often used as adjectives that describe feelings or attributes e.g.

persistant, frappant, déplaisant, contrariant, tranchant, perçant, constante.

In addition, feminine inflections of nouns less commonly used fell into this category e.g.

ourse, préfète, papesse;

also inflections related to each other such as present forms and their respective past participle forms e.g.

combine, combiné, fouille, fouillé, crève, crevé;

plural forms such as

matériaux, coordonnées, maxima, tranchées;

and singular masculine forms ending with the suffix *–u* such as

paru, résolu, retenu, conçu.

Lastly, core B2 types with FGC ranking beyond the 216,059th rank responding to criterion "**d**" (3,934 types) corresponded to about an average of 14 inflections per verbal lemma (3,244 types) such as

apparais, apparaissaient, apparaissant, apparaisse, apparaissons, apparaît, apparaîtra, apparaîtrait, apparaîtront, apparue, apparurent

derived from the 228 infinitives such as

admettre, démontrer, écouler, figurer, gêner, instruire, isoler, joindre, limiter, mener, permettre, réagir, simplifier, tarder, vaincre.

The other types found under this category were 690 words containing a majority of plural forms such as

alliances, barrières, débits, écarts, façons, gains, habitantes, images, jets, lâches, maîtrises, obscures, paniques, racines, uniformes, vainqueurs, zones;

and feminine singular forms such as

adéquate, boursière, chrétienne, définitive, éditoriale, fonctionnelle, gardienne, honteuse, immigrante, latine, massive, navale, radicale, salariée.

***Description of discrepant core B2 types*.** All B2 least discrepant types (absolute value = 7 to 99) follow, i.e.

performance, rejoindre, semestre, triple, permettre, mener, concerner, occasion, précipitation, ainsi, similaire, atteindre, aveu, indiquer, intervention, engager, dépôt, fidèle, rapport, rappel, barre, franc, placer, phase, figurer, populaire, assistance, secouer, indemnisation, stupéfiant, rejeter, définitif, réaliser, fermeté, intervenir, maritime, ralentissement, seul

in increasing order of discrepancy.

Most discrepant B2 core types in decreasing order were, i.e.,

ci, cul, équation, fonctionnel, faisceau, analogue, insignifiant, matériau, débouché, infini, défini, raccrocher, hâter, borne, épouvantable, haïr, crever, rendement, ficher, connexion, linguistique, exciter, paramètre, supplier, fondamentalement, fraction, inexistant, mépriser, aptitude, rhétorique.

Least discrepant core B2 types seemed to stem from more general notions. Most discrepant core B2 types seemed to originate from less common and more abstract concepts.

Noteworthy at this proficiency level was that, among core B2 types responding to criteria "**a**" and "**b**", the great majority of forms were nominative, adjectival, infinitive, whereas conjugated verbal inflections were rare. And, among core B2 types responding to criterion "**c**", more present and past participles were noticeable. The pool of types responding to criterion "**d**" contained mostly more frequent conjugated inflections. Additional specific notions such as music, work, the economy, hobbies and sports, ideas or disciplines were seen. Also noticeable were increasing numbers of abstract concepts and an ever increasing number of infinitives,

participial forms, familiar and even vulgar forms, less common feminine inflections and more plural inflections.

In summary, the foregoing description of core types only covered 40% of CEFR types. The following sections would determine where the remaining 60% were found, and whether they could be recommended for learning and instruction at beginning and intermediate proficiency levels.

Common only to the FDF and FGC. Of the total 369,607 unique types identified in this study, 3.1% (11,649) were types common only to the FDF and FGC. They represented 22% of FDF, and 4% of FGC types, respectively. Figure 5 depicts this second overlap section under investigation, and the first pool of frequent types to draw from for addition to core types.

Even though these types were not CEFR types, they should be given learning and teaching priority compared to types with no frequency ranking, and less qualitative importance at beginning and intermediate French language acquisition. All these types were derived from the 5,000 most frequent French lemmas. These types also had a frequency ranking in FGC. They were, therefore, all acceptable types for teaching and learning at the beginning and intermediary A1, A2, B1, and B2 levels insofar as their qualitative assessment was positive. Table 28 and 29 list inclusion and exclusion criteria, respectively, for types common only to the FDF and FGC in order of priority.

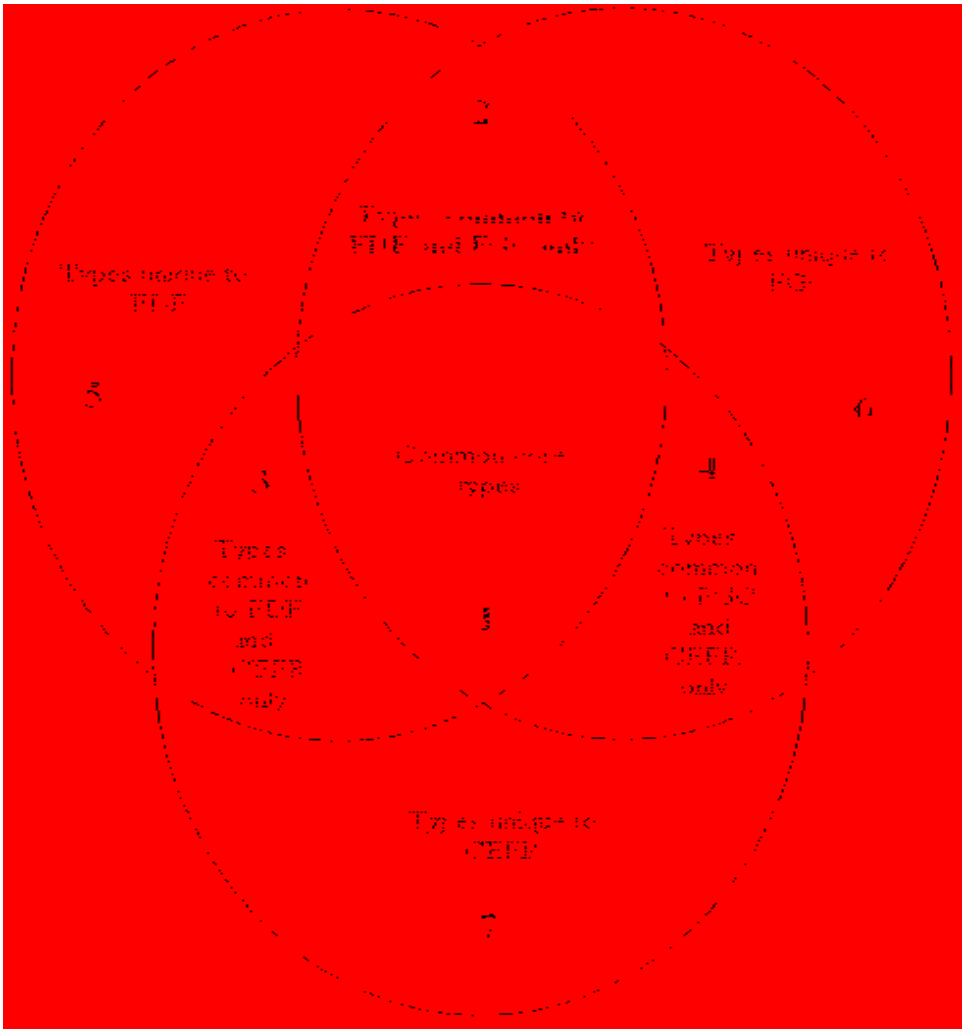

Figure 5. Types common only to the FDF and FGC

Table 28

Number of Acceptable Types Common Only to FDF and FGC by Inclusion Criteria

Inclusion Criteria	Description	Total Types
a	Δ|FDF rank -FGC rank| ≤5,000	1,570
b	FDF ranks ≤ 5,000 or FGC ranks ≤ 5,000	436
c	all inflections of FDF lemmas with positive qualitative assessment	9,181

Table 29

Number of Questionable Types Common Only to FDF and FGC by Exclusion Criteria

Exclusion Criteria	Description	Total Types
a	Rank difference ranging between 1 and 5,000 but negative qualitative assessment	0
b	FGC ranks > 5,000	11
c	Inflections of FDF lemmas but negative qualitative assessment	451

Acceptable and questionable types common only to the FDF and the FGC. In this section, the FDF-FGC ranking difference absolute value of acceptable types was less or equal to 5,000 (see Table 28 criterion "**a**"). These types were first function words, then infinitive, nominal, adjectival, adverbial forms e.g. function and pronominal words such as

dont, leur, on, jusque, parmi, outre, lequel, auprès, chez, eux, tel, lui-même, ceci;

or infinitive forms (434) such as

soutenir, conseiller, dévoiler, confier, créer, organiser, révéler, consacrer, proclamer, recommander, baptiser, couronner, témoigner, régner, œuvrer, semer, convertir, honorer, symboliser, enrichir, pardonner;

and more exclusively nominal words relative to daily affairs and themes also found in core types such as the terminology of conflict, e.g.

destruction, armement, blocage, invasion, délit, violation, complot, génocide, conquête, bombardement;

of military matters, e.g.

lieutenant, commandement, soldat, colonel, légion, déploiement;

of negociation efforts, e.g.

reconstruction, coopération, confédération, médiation, délégation, sursis, partenaire;

economic, political, or natural phenomena such as

déficit, pauvreté, rumeur, uranium, tremblement;

other human relationships and social positions not found among core types such as

reine, prince, dame, fillette, grand-mère, compatriote, chancelier, analyste, ingénieur, esclave, seigneur, prophète;

and not yet alluded to human qualities and emotions e.g.

émotion, reconnaissance, mortalité, deuil, charité, prospérité, impatience;

adverbial forms such as

quasiment, désormais, néanmoins, plutôt, provisoirement, vraiment, davantage, peut-être, guère, certes;

and nominal, verbal, or adjectival forms or a combination thereof related to the above mentioned notions such as

lutte, soviétique, assassin, hostile, mortel, tueur, préventif, homosexuel, prestigieux, héritier, raciste, menace, trône, fondateur, historique, sanglant, taxe, meurtrier, humanitaire, spirituel, terroriste, rebelle, diplomate, innocent, chiite, prématuré, précoce, intime, enthousiaste, controverse.

It was interesting to note, for instance, the forms denoting nationality or ethnicity that had not been presented in the CEFR profiles, e.g.

indien, danois, portugais, britannique, arabe, palestinien, mexicain, canadien, suédois, irlandais, marocain, serbe, kurde, syrien, néerlandais, bosniaque, égyptien, yougoslave, libanais, tchèque, basque, chiite, algérien.

Similarly, types with FDF or FGC ranks less or equal to 5,000 were also acceptable (see Table 28 criterion "**b**"). In this sub-group were found function words, e.g.

toi, moi-même, nôtre, mien, quelques-uns;

infinitive forms (103) such as

songer, dispenser, vieillir, conférer, énoncer, planifier, insérer, concilier, harceler, léguer, chier, remédier, bosser, scandaliser, revivre, rigoler;

adverbial forms e.g.

heureusement, tantôt, littéralement, dorénavant, jadis.

It was also remarkable to find vulgar or slang forms such as

foutre, ouais, mec, merde, putain, con, flic, gosse, sou, chier, farce, connerie, bordel, emmerder, salaud,

and onomatopeas such as

euh, ah, eh, hélas, hum,

all this mixed together with the divine and the devilish, e.g.

dieu, compassion, divin, pitié, majesté, doctrine, théologie, diable, pervers, fantôme, péché, démon.

What follows are more examples of the 9,181 inflections of this category with positive qualitative assessment (see Table 28 criterion "**c**"). They were any FDF/FGC type not associated to *passé simple* or exclusively subjunctive forms, i.e. not resembling indicative present inflections. Given their number, these examples were taken from types alphabetically starting with the letter *a*. For all other types of this category please refer to the electronic site http://humanities.byu.edu/frnvocab. Here again most of the inflections were formed in the same way as the following 21 related verbal inflections of the infinitive lemma *abandonner*, i.e.

abandonnaient, abandonnait, abandonnant, abandonne, abandonné, abandonnée, abandonnées, abandonnent, abandonner, abandonnera, abandonnerai, abandonneraient, abandonnerais, abandonnerait, abandonnerons, abandonneront, abandonnés, abandonnez, abandonnions, abandonnons.

Forms starting with the letter *a* deriving from 71 verbal lemmas, with an average of 13 inflections each, represent negative or extreme behavior such as

abandonner, abattre, abolir, abuser, accentuer, acharner, affaiblir, anéantir, armer, arracher, assassiner;

or more positive, protective improving behavior such as

abriter, accompagner, accroître, accueillir, adapter, ajuster, alléger, allier, aménager, ancrer, attribuer.

A number of 126 types were not verbal inflections, but rather descriptions of occupations and relationships, e.g.

actionnaire, adjointes, ambassadeurs, architectes;

origins, e.g.

albanaise, algériennes, autochtones;

qualities, e.g.

absurdes, accessible, accessoires, aérienne, agressives, alternatives, anonymes, anormale, artificiels;

processes such as

acceptation, accumulation, accusation, acquisition;

actions and their results such as

abandon, abus, acte, aboutissement, aménagements, avertissements;

plural forms of types presented in other sections such as

abris, adhésions, adieux, agressions, alarmes, albums, ambiguïtés, ambitions, âmes, atouts, avalanches.

No type whose FDF and FGC ranking difference ranged between 1 and 5,000 was found (see Table 29 criterion "**a**"). Examples of questionable FDF/FGC types whose FGC frequency ranking reached 5,000 or beyond, in increasing FGC rank order, included:

aggravant, affecté, administré, apaisant, agissant, anima, alloué, agréé, améliorant, amicale

(see Table 29 criterion "**b**").

Questionable FDF/FGC types were also inflections of 331 FDF verbal lemmas with a negative qualitative assessment (see Table 29 criterion "**c**"), i.e. less frequently used *passé simple*, or exclusively subjunctive forms such as

abstienne, agisse, caractérisèrent, confonde, convainquit, effaça, mina, pende, provins, recoure, sourisse, restreigne, soumissions.

Description of discrepant types common only to the FDF and the FGC. The 71 least discrepant types not found in the CEFR profiles also had small ranking difference ranging from 0 to 95. They were infinitive forms, i.e.

prôner, soutenir, contester, transférer, dépêcher, proposer, infliger, fuir, imputer, identifier, incarner, déployer, baser, détruire, tenter, regrouper, agir, prévenir, enfuir, convoquer, souhaiter

in increasing order of discrepancy. They were also function words, i.e.

dont, elle, lui, nous, on, je, celui, parmi

in increasing order of discrepancy. The following nominal forms also occurred, i.e.

juridiction, lutte, préjudice, texte, rendez-vous, destruction, armement, liaison, financement, affaire, finance, engagement, reconstruction, prédécesseur, vacance, humour, restructuration, territoire, crime, procédure, comédie, prélèvement, duc

in increasing order of discrepancy. Two adverbial forms were also present, i.e. *apparemment* and *quasiment*; as well as adjectival forms, i.e.

nombreux, inédit, territorial, démocratique, téléphonique, fiscal

in increasing order of discrepancy. Other forms found in least discrepant types were the preposition *parmi*, or noun/adjective forms such as

complice, coupable, général, homosexuel, nazi, partenaire, sinistre, soviétique

the noun/adverb/preposition *outre*, the preposition/adverb/conjunction *jusque*, and the determinant/adjective/pronoun *leur*. They are listed in increasing order of discrepancy in Appendix D.

The 132 most discrepant types not presented in the CEFR vocabulary profiles but present in the FDF and FGC are listed in Appendix E in decreasing order of discrepancy.

Some of these forms reflected vulgar language such as

chier, emmerder, putain, merde, salaud, bordel, connerie, con.

Other forms were onomatopoeic such as

hé, euh, hein, ha, ah, eh.

Some forms were colloquial, e.g.

ouais, camoufler, rigoler, affreux, mec, bosser, gosse, foutre, truc.

Some forms were three- to six-syllable long adverbs such as

présentement, premièrement, deuxièmement, infiniment, indépendamment, assurément, inévitablement, sincèrement, aucunement, attentivement.

Some forms referred to processes such as

spécification, distorsion, compression, interaction, soumission, planification, extraction.

The lemmatized verbal forms not already mentioned were

rayer, effrayer, relire, murmurer, écrier, blâmer, spécifier, léguer, civiliser, balayer, énoncer, omettre, repenser, englober, différencier, incomber, obséder, empresser, balancer, insérer, caresser, tisser, énumérer, déformer.

This list also contained lemmatized nominal/adjectival forms such as

chéri, vérificateur, récepteur, autochtone, terrien, primitif, insensible, insensé, idiot, cynique, contrevenant,

or adjectival forms such as

nominal, vulgaire, réputé, concurrentiel, ironique, humble, compréhensible, superficiel, rigide, convenable, avantageux, honorable, pertinent, notoire, décent, explicite.

In this FDF/FGC overlap area, a number of function words absent in the CEFR profile notions appeared. An enlarged pool of frequently used terminology related to conflict and reconciliation, military matters, negotiation, economic, political, natural phenomena, human relationships, feelings, social relations, not used in the core, was available. More infinitives, new adverbs, new nationalities, more vulgar or slang forms, onomatopeas, all mixed with theological terminology seemed to characterize this section. Words, formerly marginal and slangy, were moving front stage. All of these types seemed to play a part in describing the classic and more graphic stories of war and peace in distant lands people are getting from the news.

Common only to the CEFR and FDF. Of the total 369,607 unique types identified in the study, 3.2% (12,056)) were types common only to the CEFR profiles and the FDF. They represented 22.5% of FDF types. Figure 6 depicts the third overlap section under investigation, and the second pool of frequent types to draw from for inclusion to the core CEFR types. These types represented close to a third (29.2%) of the total CEFR profile types. They were characterized by their only relation to the FDF (see Table 30), and not to the FGC. Table 31 shows the breakdown of acceptable CEFR/FDF types which does not increase smoothly in numbers. B1 and B2 types were respectively three times and four times as numerous as A1 types of this category, whereas A2 types were barely 29% more numerous than A1 types. Table 32 shows the breakdown of questionable CEFR/FDF types which happened to amount to over half the number of types in this pool.

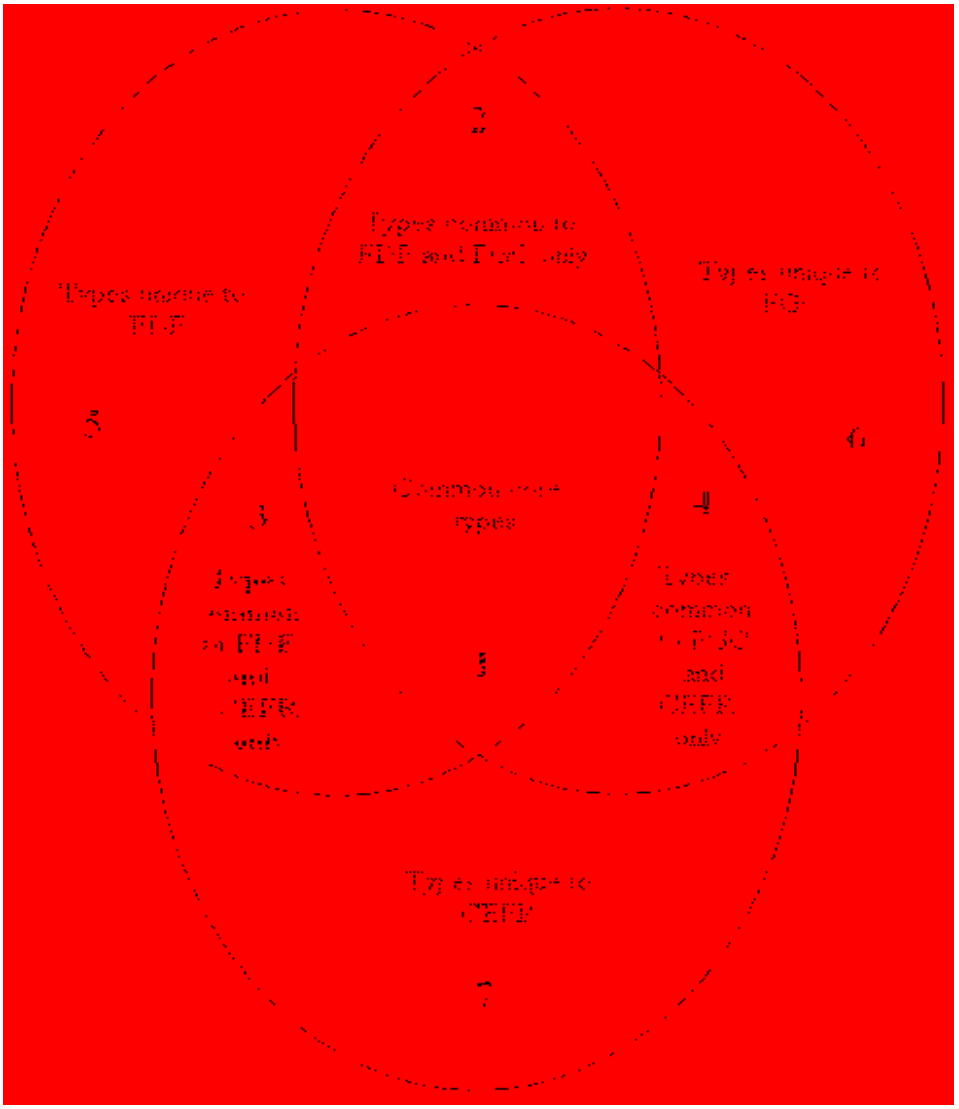

Figure 6. Types common only to the CEFR and the FDF

Table 30

Number of Types Common Only to the CEFR and FDF by Proficiency Levels

		Common Only to CEFR and FDF	
Proficiency level	CEFR Types	Number	Percentage
A1	5,853	1,613	27.6
A2	5,234	1,816	34.7
B1	11,913	3,865	32.4
B2	18,255	4,762	26.1
Total	41,255	12,056	29.2

Table 31

Number of Acceptable Types Common Only to the CEFR and FDF by Inclusion Criterion

Inclusion Criterion	Description	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
a	Inflections of FDF lemmas with positive qualitative assessment	598	774	1,838	2,452	5,663

Table 32

Number of Questionable Types Common Only to the CEFR and FDF by Exclusion Criterion

Exclusion Criterion	Description	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
a	Inflections of FDF lemmas but negative qualitative assessment	1,015	1,042	2,027	2,310	6,393

Acceptable CEFR/FDF A1 types (see Table 31 criterion "**a**") were mainly forms of the French *présent, imparfait, futur,* or *conditionnel* such as

achèteras, achèteriez, achètes, achetiez

of common 92 verbal lemmas such as

acheter, boire, chanter, danser, écouter, fermer, habiter, inviter, jouer, lire, manger, naître, ouvrir, parler, raconter, servir, téléphoner, vivre;

a few present or past participle forms such as

sortissant, *continués, coûtée, coûtées, répondue;*

as well as a few plural forms such as

juins, bonsoirs, mesdemoiselles, minuits, soifs.

So were acceptable CEFR/FDF A2 types (see Table 31 criterion "**a**") such as

allumerai, allumerais, allumerez, allumiez

of 99 common verbal lemmas such as

amuser, bouger, calculer, déclarer, employer, guérir, habiller, inscrire, jeter, marier, opérer, porter, refuser, sauver, traduire, vérifier;

and a few plural forms e.g.

doctoresses, vitæ, vouloirs;

and likewise, acceptable CEFR/FDF B1 and B2 types (see Table 31 criterion "**a**") of 194 and 224 common and frequent verbal lemmas, respectively, such as

adopter, composer, délivrer, diviser, élever, fier, généraliser, interpréter, loger, multiplier, nommer, ordonner, prier, quitter, ralentir, séduire, taire, unir, verser

for B1 types, and such as

accorder, bouleverser, céder, débuter, éclaircir, fabriquer, haïre, incliner, jurer, livrer, minimiser, nier, paraître, ramener, satisfaire, tarder, vaincre

for B2 types.

Regarding questionable CEFR/FDF types at level A1, A2, B1, and B2 (see Table 32 criterion "**a**"), forms identified were exclusively an average of 10 *passé simple* or exclusive subjunctive forms per verbal lemma such as

achetai, achetâmes, achetas, achetasse, achetassions, achetâtes

of respectively 105, 98, 191, and 214 frequent infinitives. CEFR/FDF A1 types looked like the following examples, e.g.

aimer, boire, comprendre, décider, écrire, finir, gagner, habiter, inviter, jouer, laver, mesurer, naître, ouvrir, parler, recevoir, sentir, tenir, voir.

CEFR/FDF A2 types yielded inflections derived from infinitives such as

accepter, bouger, calculer, déclarer, employer, grossir, inscrire, mélanger, oublier, pleurer, remplir, sauver, traduire, voler;

and the three unusual forms

drôlesse, drôlesses, porters.

CEFR/FDF B1 types also yielded similar inflections from infinitives such as

apprécier, approuver, causer, demeurer, ennuyer, fonder, gouverner, interroger, mentir, nourrir, observer, piloter, questionner, quitter, rechercher, saisir, transformer, unir.

And, so did CEFR/FDF B1 types such as

accorder, briser, combiner, décéder, engendrer, fabriquer, grimper, hisser, invoquer, limiter, mener, orienter, pénétrer, raccrocher, secouer, tarder, vaincre, vanter, viser.

In this section, CEFR types were common to types derived from the 5,000 most frequent French lemmas. It was no surprise to witness an important share of the inflections of most frequent FDF verbal lemmas. The remarkable observation was that CEFR/FDF types were mainly derived from verbal forms.

Common only to the CEFR and FGC. Types common only to the CEFR and FGC represented 14.1% (5,817) of the CEFR total (see Table 33 for a breakdown by proficiency level). Figure 7 depicts the fourth overlap section under investigation, and the third pool of types to draw from for inclusion in the core CEFR types. Here, CEFR types were characterized by their commonality with frequent FGC types. As with the prior sections, inclusion criteria defined acceptable types (see Table 34). They, in turn, led to the definition of ordered criteria for exclusion from the CEFR profiles (see Table 35).

Table 33

Number of Types Common Only to the CEFR and FGC by Proficiency Levels

		Common Only to CEFR and FGC	
Proficiency Level	CEFR Types	Number	Percentage
A1	5,853	255	4.4
A2	5,234	406	7.8
B1	11,913	1,249	10.5
B2	18,255	3,907	21.4
Total	41,255	5,817	14.1

Table 34

Number of Acceptable Types Common Only to the CEFR and FGC by Inclusion Criteria

Inclusion Criteria	Description	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
a	FGC ranks ≤ 5,000	10	18	19	56	103
b	FGC ranks > 5,000 but positive qualitative assessment	107	135	419	1,465	2,126

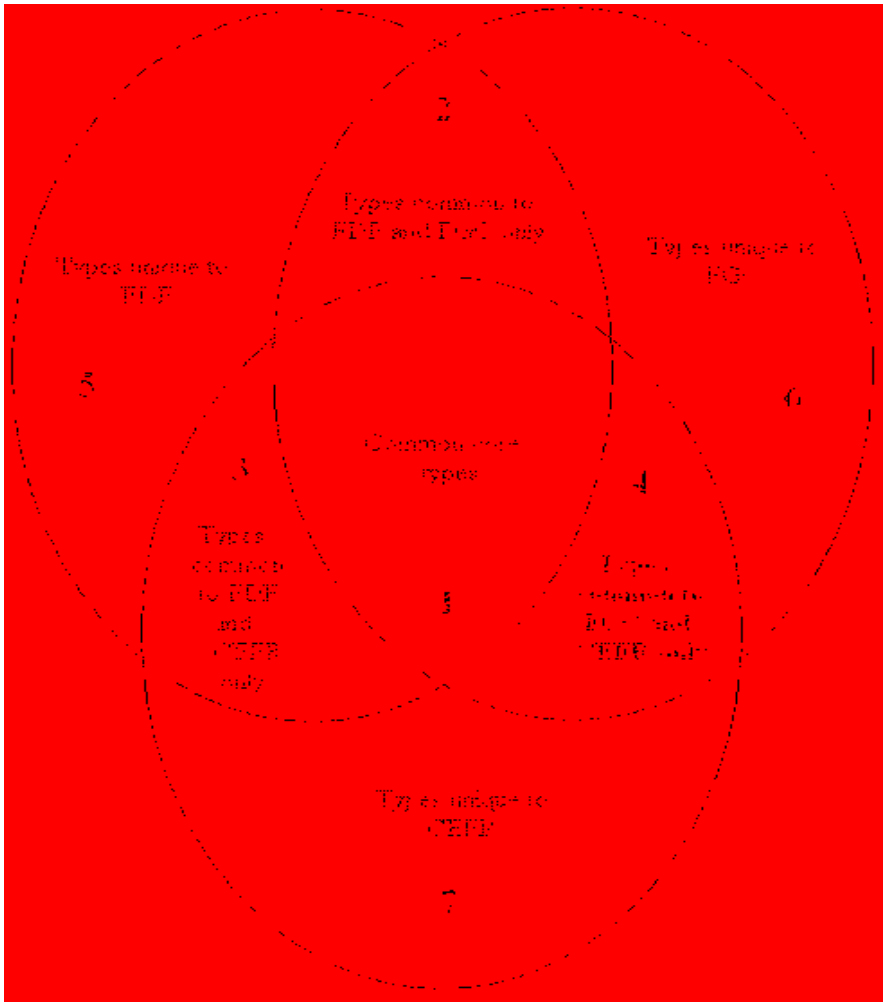

Figure 7. Types common only to the CEFR and FGC

Table 35

Number of Questionable Types Common Only to the CEFR and FGC by Exclusion Criteria

Exclusion Criteria	Description	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
a	FGC ranks > 5,000 and negative qualitative assessment	0	0	3	4	7
b	No FGC ranks or FGC ranks > 197,914	138	253	808	2,382	3,581

Acceptable CEFR/FGC A1 types (see Table 34 criterion "**a**") with a FGC rank less or equal to 5,000 were as follows

au, parce, rugby, poivre, salade, moto, centimètre, pluriel, jus, parking

in increasing FGC rank order. Acceptable CEFR/FGC A1 types meeting criterion "**b**" (see Table 34) were types with FGC ranks higher than 5,000 but positive qualitative assessment; they happened again to be very common French words, necessary for everyday life, e.g.

riz, gramme, internet, supermarché, locataire, lapin, camping, banane, cuillère, yaourt, boulangerie, dentiste, tasse, jupe, laine, aspirine, chaussette, stylo, savon, crayon, pull, mél

in increasing FGC rank order. Acceptable CEFR/FGC A2 types (see Table 34 criterion "**a**") were types with FGC ranks less or equal to 5,000, i.e.

du, préfecture, la, arrondissement, grippe, maillot, mosquée, pelouse, sida, natation, sauce, canton, vaccin, nager, ambulance, cathédrale, dauphin, basket

in increasing FGC rank order. Acceptable CEFR/FGC A2 types meeting criterion "**b**" were common and useful types with FGC ranks above 5,000 but with positive qualitative assessment (see Table 34 criterion "**b**"), such as

discothèque, synagogue, antibiotique, saumon, jambon, haricot, purée, crevette, sirop, thon, rôti, pizza, pâté, marron, crêpe, sorbet, apéritif, pâtisserie, sandwich, couscous, brioche

in increasing order of FGC ranking. Acceptable CEFR/FGC B1 types were types with FGC ranks less or equal to 5,000 (see Table 34 criterion "**a**") e.g.

nuageux, hospitaliser, ensoleillé, bronze, autocar, four, compagne, intérim, respiratoire, documentaire, fracture, stationner, valider, orthodoxe, secourir, adhérent, neuvième,

in increasing rank order; or with FGC ranks beyond 5,000 and a positive qualitative assessment (see Table 34 criterion "**b**"). Examples chosen were taken from the medical repertoire, e.g.

chirurgical, hospitalisation, généraliste, chirurgie, vitamine, vacciner, entorse, contagieux, pédiatre, digestif, allergique, anesthésie, rhume, thermomètre, asthmatique, diététique, radiologique, inflammatoire, pansement, digestion, vacciné, hygiénique, varicelle, compresse, jaunisse, samu,

in increasing rank order. Acceptable CEFR/FGC B2 types were types with FGC ranks less than or equal to 5,000 (see Table 34 criterion "**a**"). Examples here were taken from sports terms, e.g.

coureur, cycliste, athlétisme, rallye, finaliste, skieur, footballeur

in increasing rank order. Acceptable B2 examples corresponding to criterion "**b**", in increasing rank order, were types with FGC ranks beyond 5,000 and a positive qualitative assessment (see Table 34) such as these taken from geography terminology, e.g.

littoral, péninsule, côtier, méditerranéen, pic, mont, plaine, torrentiel, hémisphère, insulaire, delta, amont, superficie, torrent, rivage, prairie, sentier, fluvial, falaise, estuaire, embouchure, océanique.

No example of questionable types corresponding to criterion "**a**" was found (see Table 35). Questionable CEFR/FGC A1 types (see Table 35 criterion "**b**") were types with no FGC rank or ranks beyond 197,914 such as plural forms of everyday life words for clothing, grooming, food, appliances, locations, objects, e.g.

aspirines, baguettes, carrefours, dentifrices, embarquements, gants, jupes, pharmacies, réfrigérateurs, serviettes, tickets, virgules, yaourts.

Similarly, no example of CEFR/FGC A2 types meeting criterion "**a**" was found. Examples of questionable CEFR/FGC A2 types with no FGC ranks or ranks beyond 197,914 (see Table 35 criterion "**b**") were again mostly plural forms of common words such as

bottes, cachets, écharpes, frites, guitares, haricots, infirmeries, jobs, lessives, maillots, pâques, ramadans, synagogues, tulipes, vaccins.

Examples of questionable CEFR/FGC B1 types with no FGC rank or ranks beyond 197,914 (see Table 35 criterion "**b**") were taken from zoology terminology, and were noticeably plural forms in general, e.g.

araignées, coqs, coquillages, crabes, crocodiles, faunes, femelles, huîtres, mâles, tigres, tigresses, tourteaux, zoos.

Finally, examples of questionable CEFR/FGC B2 types with no FGC rank or ranks beyond 197,914 (see Table 34 criterion "**b**") are used in music terminology e.g.

accompagnateur, accordéon, accordéoniste, choeurs, chorales, clarinettes, solistes, sonorités.

This section characterized the partial CEFR/FGC overlap. Very useful types, sometimes described as "available" types were present here. They belonged to general and, more heavily, to the specific CEFR notions in categories such as quantity and size, foods, sports, modern communication, services (e.g. medical), zoology, and geography. More questionable, because less frequent, were the plural inflections of these forms.

Unique to the FDF. Of the 53,511 types counted in the FDF, about one out of four (13,157 types, 24.6%) was unique to the FDF. Being part of the FDF, the great majority of these types were derived from the 5,000 most frequent French lemmas. They could not be found in the CEFR profiles or the FGC. BDLex having been used to obtain these types, no frequency and rank order data was available for them. Figure 8 depicts this fifth section under investigation, and the fourth pool of types to draw from for addition to core CEFR types.

First, plural forms, mainly feminine, were noticeable. They referred to feminine qualities or occupations, e.g.

chancelières, colonelles, comtesses, diablesses, entrepreneuses, historiennes, poétesses;

to virtues not often used in the plural form, e.g.

compassions, innocences, intégrités, intimités, patiences, pitiés;

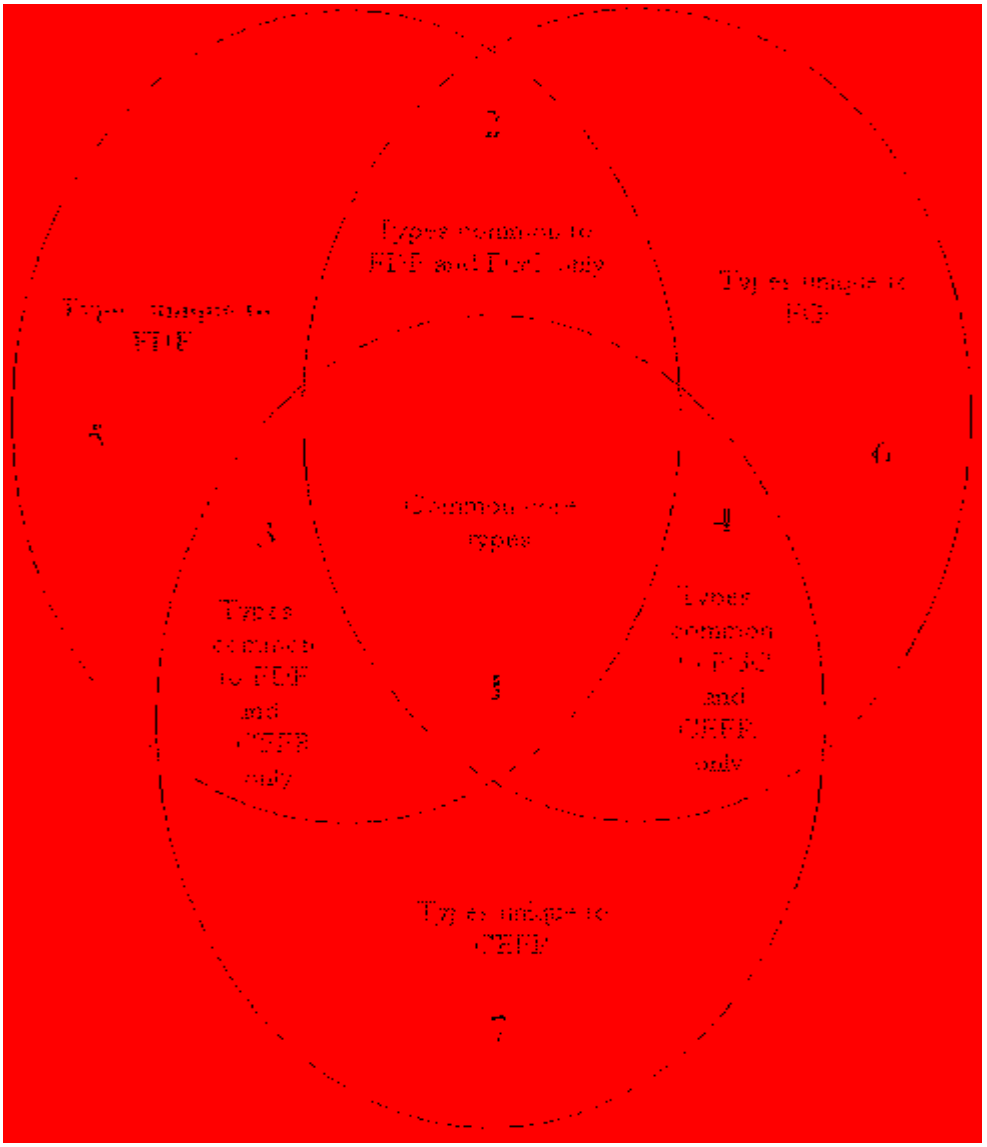

Figure 8. Types unique to the FDF

or to forms usually used in the singular e.g.

compétitivités, envergures, mondialisations, prostitutions, vigueurs.

Masculine plural forms usually seen in the singular also belonged to this group, e.g.

néants, renouveaux, essors, humours, isolements, uraniums.

In addition, plural and/or singular forms recognizable because of their *œ* ligature belonged to this FDF section, i.e.

bœuf, bœufs, cœur, cœurs, main-d'œuvre, manœuvre, manœuvres, mœurs, nœud, nœuds, œil, œuf, œufs, œuvre, œuvres, sœur, sœurs, vœu, vœux.

Most noticeable however were the 13,087 (99.5%) the inflections of 575 frequent verbal forms with an inflection count per verbal form ranging from 1 to 44, e.g., the one form *œuvrer* and 44 inflections for *comparaître*:

comparai, comparaissais, comparaisses, comparaissez, comparaissiez, comparaissions, comparaissons, comparaîtrai, comparaîtrais, comparaîtras, comparaîtrez, comparaîtriez, comparaîtrions, comparaîtrons, comparâmes, comparas, comparasse, comparassent, comparasses, comparassiez, comparassions, comparât, comparâtes, comparerai, compareraient, comparerait, compareras, comparèrent, comparerez, compareriez, comparerions, comparerons, compariez, comparions, comparûmes, comparurent, comparus, comparusse, comparussent, comparusses, comparussiez, comparussions, comparût, comparûtes.

These forms are listed in increasing order of number of inflections, i.e., 1 to 9 inflections each for 7 verbal lemmas (followed by number of inflections), e.g.

commander (1), carburer (2), ensuivre (5), agir (7);

10 to 19 inflections each for 113 verbal lemmas, e.g.

réjouir (10), accomplir (11), bâtir (12), souhaiter (13), considérer (14), accueillir (15), applaudir (16), abandonner (17), accompagner (18), améliorer (19);

20 to 29 inflections each for 422 verbal lemmas, e.g.

survivre (20), accentuer (21), abriter (22), accumuler (23), chiffrer (24), dater (25), enchaîner (26), handicaper (27), insérer (28), moderniser (29);

30 to 39 inflections each for 30 verbal lemmas, e.g.

conquérir (30), repenser (31), industrialiser (32), civiliser (34), vêtir (35);

40 to 44 inflections each for 3 verbal lemmas, i.e.

rayer (40), effrayer (41), comparaître (44).

All types unique to FDF were acceptable for instruction before reaching the intermediary-advanced threshold on the basis that they were derived from the most frequent French forms (see Table 36). However it is clear from this analysis that they were not as frequent as the forms found in the common core, or the overlap of the FDF with the FGC or the CEFR. Thus, the latter should take precedence over types unique to the FDF.

Verbal inflections which were more numerous within this FDF subgroup indicated that the forms were less frequent. This is why the FDF forms of this section should be taught by increasing number of inflections (see Table 37). Acceptable and questionable types unique to the FDF both reached over the amount of 6,000 types. Acceptable types only outnumbered questionable types by 703 types.

Table 36

Number of Acceptable Types Unique to FDF by Inclusion Criterion

Inclusion Criterion	Description	Total Types
a	Frequency ranks from 1-5,000 and positive qualitative assessment	6,930

Table 37

Number of Questionable Types Unique to FDF by Exclusion Criterion

Exclusion Criterion	Description	Total Types
a	Frequency ranks from 1-5,000 but negative qualitative assessment	6,227

Acceptable and questionable types unique to the FDF were taken from the same pool of inflections and derived from verbal FDF lemmas. Their description will be limited to a few pertinent examples. For instance, the verbal lemma *abandonner* (abandon) numbered a total of 17 inflected types unique to the FDF. Of these 17 inflections, seven forms, i.e.

abandonnais, abandonneras, abandonnerez, abandonneriez, abandonnerions, abandonnes, abandonniez

were considered acceptable, and 10 forms, i.e.

abandonnai, abandonnâmes, abandonnas, abandonnasse, abandonnassent, abandonnasses, abandonnassiez, abandonnassions, abandonnât, abandonnâtes

were considered questionable. The verbal lemma *reconstruire* (reconstruct) numbered a total of 22 inflected types unique to the FDF. Of these 22 inflections, 10 forms, i.e.

reconstruiraient, reconstruirais, reconstruiras, reconstruirez, reconstruiriez, reconstruirions, reconstruis, reconstruisais, reconstruisez, reconstruisiez

were acceptable, and 12 forms, i.e.

reconstruises, reconstruisîmes, reconstruisirent, reconstruisis, reconstruisisse, reconstruisissent, reconstruisisses, reconstruisissiez, reconstruisissions, reconstruisit, reconstruisît, reconstruisîtes

were questionable. Likewise, the verbal lemma *relire* (read again, reread) numbered a total of 30 inflected types unique to the FDF. Of these 30 inflections, 19 forms, i.e.

relira, relirai, reliraient, relirais, relirait, reliras, relirez, reliriez, relirions, relirons, reliront, relis, relisaient, relisais, relisait, relisiez, relisions, relisons, relues

were acceptable, and 11 forms, i.e.

relises, relûmes, relurent, relusse, relussent, relusses, relussiez, relussions, relut, relût, relûtes

were questionable.

In this section of types unique to the FDF, forms, typically feminine plural in nature, were emerging. Also noticeable were masculine plural forms usually seen in the singular. Striking, and possibly due to electronic computing, all regular inflections of types with "*œ*"

ligature were concentrated here. But most importantly, varying numbers of inflections derived from very frequent verbal lemmas were based in this portion of lexical data.

Unique to the FGC. Eighty-two percent (i.e. 303,546 types) of this evaluation's grand total were unique to the FGC. The initial count of close to two million was narrowed to 337,661 types by adopting a threshold of 10 occurrences or more for inclusion in the study. Figure 9 depicts this sixth section under investigation, and the fifth and last pool to draw from for addition to the common core. Types unique to the FGC reflected terminology found in the French newswire register which is not found elsewhere.

Table 38 describes criteria used to define acceptable types unique to FGC. Table 39 conveys criteria used to distinguish questionable types unique to FGC.

Table 38

Number of Acceptable Types Unique to FGC by Inclusion Criteria

Inclusion Criteria	Description	Total Types
a	Frequency ranks from 1-5,000 and positive qualitative assessment	233
b	FGC frequency ranks > 5,000 but positive qualitative assessment	745

Table 39

Number of Questionable Types Unique to FGC by Exclusion Criteria

Exclusion Criteria	Description	Total Types
a	FGC frequency ranks ≤ 5,000 but negative qualitative assessment	1,345
b	FGC frequency ranks > 5,000 and negative qualitative assessment	301,223

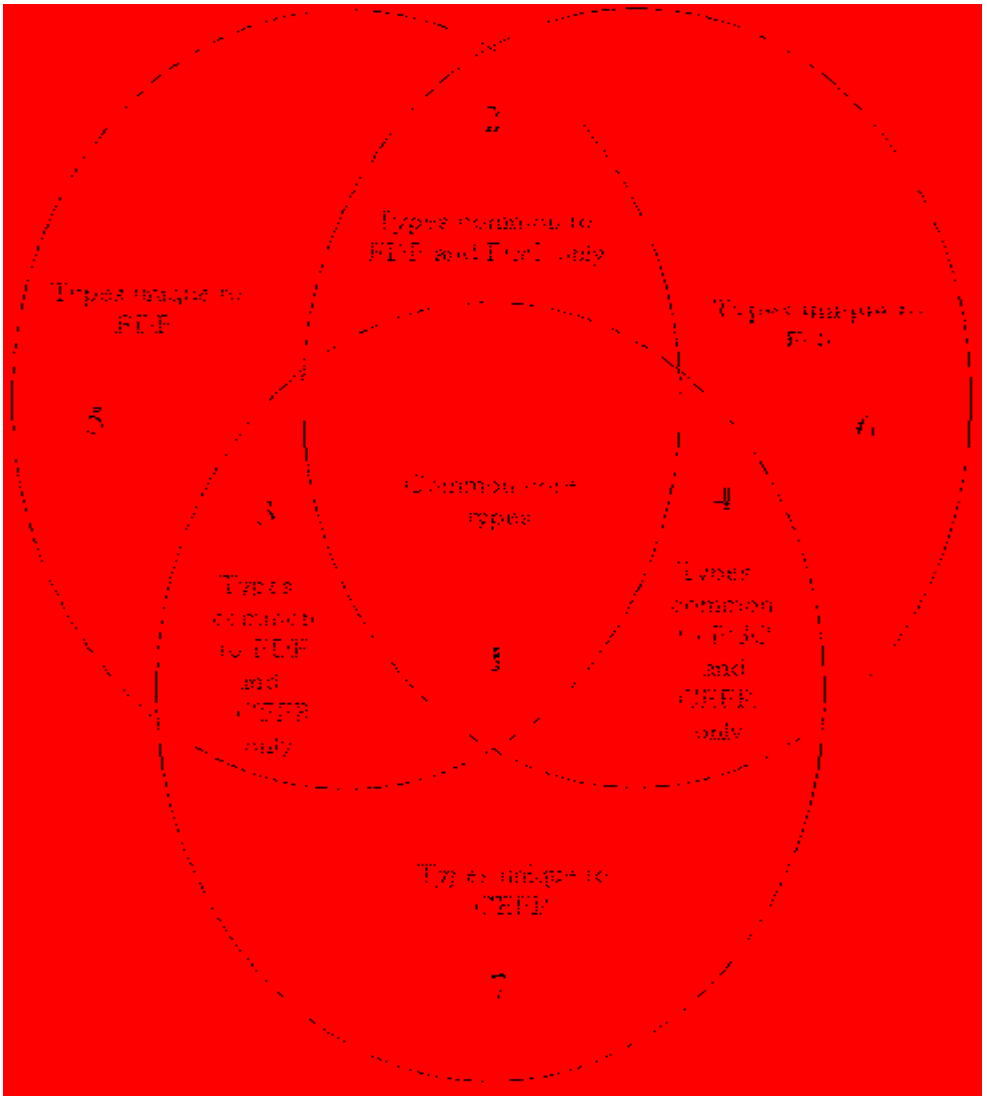

Figure 9. Types unique to the FGC

Types unique to FGC with a frequency rank ranging between 1 and 5,000 and a positive qualitative assessment were acceptable for inclusion in the CEFR profiles (see Table 38 criterion "**a**"). Examples of these acceptable FGC types come from military terminology e.g.

agresser, antiterroriste, arsenal, artillerie, assaillant, bastion, bouclier, caserne, duel, embuscade, émissaire, emprisonnement, encercler, espionnage, fusillade, grenade, guérilla, insurgé, insurrection, milicien, obus, offensif, patrouille, rescapé, riposter, roquette.

Types unique to FGC with a rank beyond 5,000 but with a positive qualitative assessment should also be considered as acceptable to inclusion in the CEFR profiles (see Table 38 criterion "**b**"). Examples of acceptable types in this category were chosen from nationalities, ethnicities, or corresponding languages not yet included in other sections, e.g.

albanophone, angolais, bolivien, cambodgien, catalan, centrafricain, chypriote, colombienne, coréen, dominicain, équatorien, éthiopien, finlandais, flamand, haïtien, helvétique, hindou, hollandais, libérien, lituanien, maghrébin, nippon, nordique, nordiste, normand, péruvien, phocéen, sénégalais, tchadien, thaïlandais, tibétain, vietnamien.

Questionable FGC types manifest here were commonly used capitalized acronyms such as

DVD, ADN, EDF, TGV, TVA, URSS.

Hyphenated FGC types occurred as well such as

attentat-suicide, c'est-à-dire, couvre-feu, demi-finale, ex-Yougoslavie, sans-papiers, raz-de-marée.

And so did FGC types with apostrophes such as

d'abord, d'accord, d'ailleurs;

proper names, e.g.

Albright, Alliot-Marie, Armstrong;

or capitalized types such as

COMPTE-RENDU, HEBDOMADAIRE, METEO,

since all these features went against our definition of types.

Questionable FGC types with FGC ranks beyond 5,000 but with negative qualitative assessment (see Table 39 criterion "**b**"), starting with the letters k and l, were less frequent, more technical or specialized words such as

kabbaliste, kabyle, kalachnikov, kaléïdoscope, kamikaze, kangourou, keynésien, khmère, kinésiologie, kinétoscope, labelliser, labiale, labyrinthe, lacération, laconisme, lacrymal, lacustre, lambeau, lancinant, langoureuse, laps, latence, laudateur, laxisme, lénifiant, lésiner, lézarder, libidineux, lipidique, loquace, loufoque, louvoiement.

Other examples of questionable types unique to the FGC were misspelt words such as

pèsant, pesera, pésera, péserait, *pessismisme, petanque, phénoméne, phènomène, phenomène, phenomene*;

English types such as

memorycard, merchandising, *monkeypox*;

and types from other languages such as Arabic,

mouhamoud, mostafa, moudjahdine,

German,

leiden, leider, leicht, klang,

Italian,

mezzogiorno, mezzo-soprano, *montegiordano*,

Japanese,

sayonara, sayoko, takimoto,

or Spanish,

muchas, muchachos, muerta.

Also evident were first, middle, and surnames spelled different ways, some of which have become brand names, i.e.

mercedes, mercedès, mercédes, mercédès;

specialized medical, chemical, and biological types such as

méningiome, méthyléthylcétone, méthyltestostérone, monoclonaux, mycobactérie, myocardiopathie,

or yet specialized names of plants or insects such as

millepertuis, mille-pattes, *mygales*.

Types unique to FGC were colored by notions evoked in newswire text and not yet revealed in prior sections. They dealt, for instance, with commentaries on conflicts in various lands, thus ensuing types related to these topics. They also dealt with advanced and specialized scientific topics of a variety of disciplines. Noted were also acronyms, spelling errors, and non-French forms.

Unique to the CEFR. Of the total 41,255 types found in CEFR vocabulary profiles, 16.3% (6,733) were unique to the profiles (see Table 40). Figure 10 depicts the seventh section under investigation, and the main pool for exclusion from the CEFR profiles.

Table 40

Number of Types Unique to the CEFR Profiles by Proficiency Levels

		Unique to CEFR Profiles	
Proficiency level	Total CEFR Types	Number	Percentage
A1	5,853	126	2.2
A2	5,234	354	6.8
B1	11,913	1,585	13.3
B2	18,255	4,668	25.6
Total	41,255	6,733	16.3

All types unique to the CEFR profiles are questionable for instruction before reaching the intermediate-advanced threshold. By definition they are not frequent. However, some inflections of the lemmas they derive from are more frequent than they are. These more frequent CEFR forms were found in the common core or overlapping with the FDF or FGC. Thus, verbal inflections found in overlapping sections should take precedence over types unique to the CEFR profiles. Verbs whose inflections were more numerous in this section demonstrated that some of

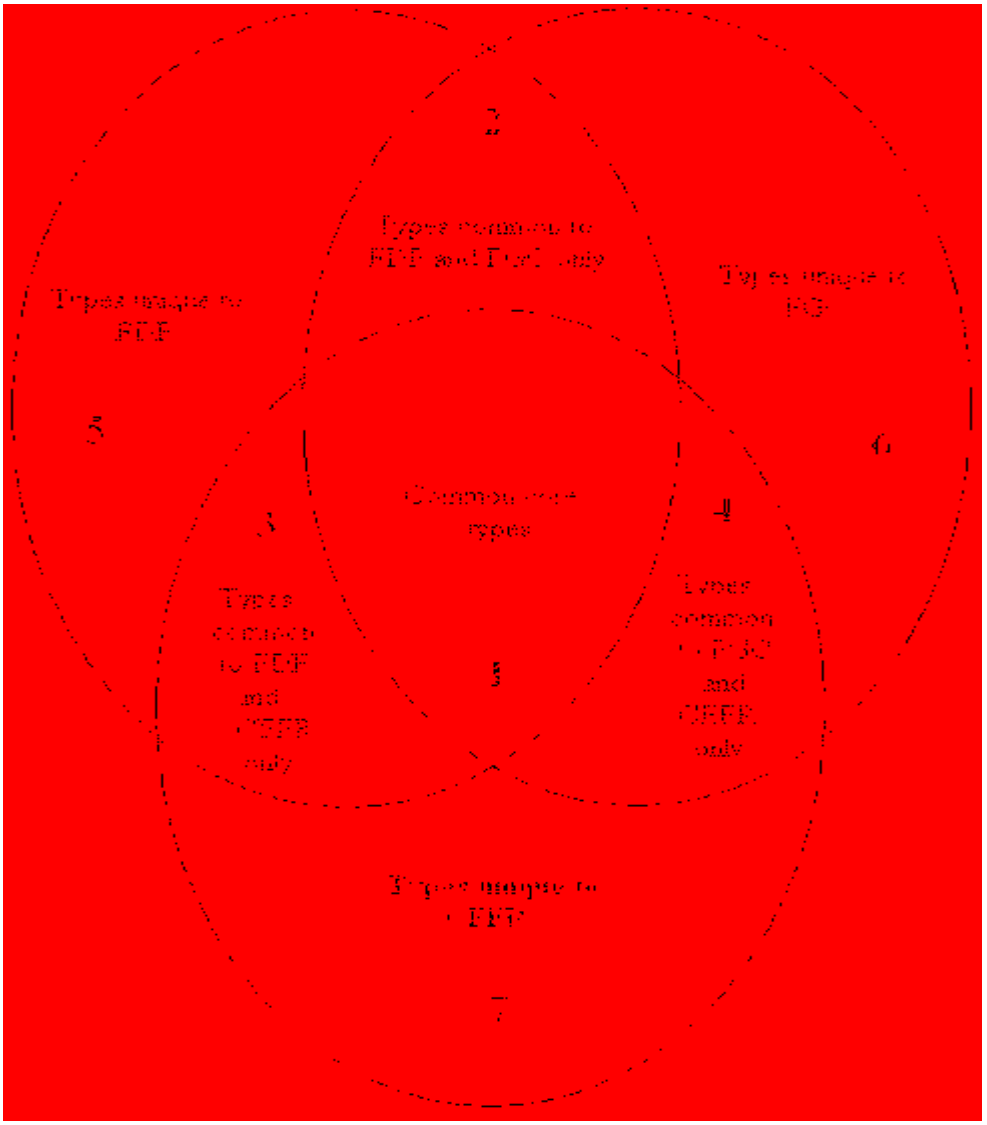

Figure 10. Types unique to the CEFR

their forms were less frequently used. Certainly, a number of types of this section could receive a positive qualitative assessment since they derived from useful infinitive forms in given contexts such as

coiffer, brosser, doucher, épeler

at level A1 with 13, 15, 16, and 23 inflections respectively; or

camper, skier, cuisiner, tousser

at level A2 with 12, 13, 15, and 17 inflections respectively; and

bavarder, saler, poivrer, sucrer

at level B1 with 16, 17, 18 and 19 inflections respectively. They could also receive a positive qualitative assessment when they were not *passé simple* or exclusively subjunctive types such as

balbutiâmes, anesthésiassiez, boxât,

which are used infrequently, or when they did not contain forms hardly ever used in the plural such as

méditerranées, christianismes, japons.

However, the considerable number of types found and analyzed in earlier sections do not justify including types unique to the CEFR profiles. Consequently, types unique to the CEFR profiles should only be added in instructional materials after inclusion of all other acceptable types of each preceding section. Illustrative examples of these questionable types are now provided.

Questionable A1 types here were feminine forms such as

adjectives, agnelles, écologies, grammaires, validités;

plural forms of foreign types with variable spelling, e.g.

babys, foots, rugbys;

and *passé simple* and subjunctive forms, in total 11, 12, 12, and 12 forms respectively, connected to the infinitive forms *brosser, coiffer, doucher,* and *épeler* such as

brossâmes, brossassiez, brossât, coiffâmes, coiffassiez, coiffât, douchâmes, douchassiez, douchât, épelâmes, épelassiez, épelât

(see Table 40).Questionable A2 types were also mostly feminine inflections, singular and plural, such as

aspiratrice, aspiratrices, *expéditrice, expéditrices, mécaniciennes, natations, pâtissières, poissonnières, radiologies, répondeuse, répondeuses;*

and verbal inflections of the following lemmas (followed by number of inflections):

balader (11), boulanger (37), camper (10), cocher (12), cuisiner (12), débrancher (12), divorcer (11), maigrir (8), nager (12), parfumer (12), skier (12), tousser (12),

e.g.

baladassiez, boulangeassent, campâtes, cocha, cuisinassent, débranchas, divorçâmes, maigrîtes, nageassions, parfumasse, skiâtes, toussassiez

(see Table 40). Questionable B1 types were essentially singular forms such as

audibilité, *campeuse, pacsée, royalisme, vidéoprojecteur*;

plural forms such as

anciennetés, bries, crémeries, disjonctions, *équateurs, faunesses, gazoles, honnêtetés, jardinages, langagières, médiathèques, pessimismes, sexualités, varicelles*;

and *passé simple* or exclusively subjunctive inflections such as

accouchai, accouchâmes, accouchas, accouchasse, accouchassent, accouchasses, accouchassiez, accouchassions, accouchât, accouchâtes, accouchèrent

of the following infinitive forms, e.g.

blanchir (5), composter (13), dégoûter (12), éduquer (12), fiancer (11), garer (8), hospitaliser (12), jardiner (12), lessiver (12), meugler (36), noircir (5), oranger (32), parquer (13), réchauffer (12), saler (10), vacciner (12)

(see Table 40). Finally, questionable B2 types were predominantly verbal inflections (92% or 4,270 types) of 143 infinitive forms with an average of 30 inflections per verb ranging from 1 to

48. This great quantity of verbal inflections revealed a considerable number of infrequently used forms. Examples of verbal infinitive forms with their increasing number of inflections in parentheses follow; for instance, 1 to 19 inflections each for four verbal lemmas, i.e.,

cokoter (1), bruiner (14),pluviner (15), assombrir (17);

20 to 29 inflections each for 58 verbal lemmas, e.g.

engloutir (21), redoubler (22), vomir (23), bouder (24), élancer, (25), frotter (26), tremper (27), agenouiller (28), légiférer (29);

30 to 39 inflections each for 80 verbal lemmas, e.g.

aveugler (30), tordre (31), vexer (32), naturaliser (33), trotter (34), vidanger (35), épiler (36), engueuler (37), désodoriser (38), conceptualiser (39);

and one outlier *bégayer (48)*. These verbal types covered activities of physical/personal/domestic care such as

épiler, désodoriser, teindre, colorer, brunir, frotter, tremper, héberger, humidifier.

They also referred to sounds such as

assourdir, grincer,

odors such as

cokoter, empester,

or to changes such as

dérouter, révolutionner, décommander, disjoindre,

or comparisons such as

égaler, surpasser.

Many of these types are used in culinary or chemistry terminology such as

déguster, rôtir, savourer, cristalliser, diluer, mariner, moudre, assaisonner, mijoter, beurrer, écailler, tartiner, graisser, surgeler.

Others referred to ways of eating such as

engloutir, dévorer,

or to ways of talking such as

prêcher, calomnier, injurier, grogner, bafouiller, chuchoter, déclamer, dénommer, gronder, médire, balbutier, gueuler, bégayer.

Others still referred to ways of thinking such as

postuler, surestimer, synthétiser, discerner, théoriser, conceptualiser.

Some distinguish various motions and movements such as

élancer, fouler, mouvoir, escalader, distancer, ramper, déguerpir, foncer, longer, dévaler, boiter, décamper, détaler, trotter, *accroupir, agenouiller, courber*

or ways of touching such

cueillir, tordre, tâter, palper.

Some verbal types alluded to group and legal activities such as

assembler, affilier, légiférer, naturaliser,

or financial/accounting matters such as

déprécier, cotiser, dénombrer, décroître,

or health matters such as

anesthésier, vomir, déprimer, panser, ausculter, péter.

Some described feelings such as

bouder, irriter, envier, vexer, réconforter, agacer, déconcerter, embêter, froisser, apprivoiser, chagriner, taquiner, brusquer.

Some types referred to the school system such as

redoubler, recaler.

Others described the weather such as

bruiner, pluviner.

Some dealt with vision such as

assombrir, aveugler, loucher;

some with work related to vehicles such as

remorquer, vidanger.

A good number also dealt with ways of organizing such as

tasser, emballer, entasser, empiler, éparpiller, déssécher, synchroniser.

The remaining forms were plural forms (6%) and singular forms (2%). Examples of the 295 plural forms were taken from the food terminology, e.g.

avoines, bouillabaisses, cannelles, gigots, jambonneaux, macédoines, mâches, muscats, paprikas, persils, seigles, vacherins.

Examples of 108 singular forms were drawn from mathematics terminology, e.g.

abscisse, algébrique, bissectrice, commutativité, équidistant, inductif, isocèle, ordinal, polynôme, vectrice

(see Table 40).

Types unique to the CEFR profiles comprised feminine plural forms, plurals of foreign types, and passé simple or exclusively subjunctive forms. These forms are hardly ever used. They were made apparent via the BDLex methodology used to produce all possible French inflected types for this study. Types unique to the CEFR profiles also contained newer or antiquated singular forms. They reflected perhaps the propensity to use favorite words for instruction such as sounds produced by animals, e.g. *aboyer, miauler, meugler*, or think that *passé simple* forms will be useful for the reading of fiction and literary works.

Note: Thirteen B2 types unique to the CEFR profiles, i.e.

conjoindre, excepter, surgeler, graduer, côtelette, ordinal, vénérien, centilitre, empester, bruiner, équidistant, chicon, crémier,

and one B1 type

jaunissant,

ended up having a FGC ranking which they should not have had. The situation arose from the use of a lemmatized file with FGC rankings which attributed ranks to FGC inflected forms. For instance, the singular form *centilitre* even though it was only found among types unique to the CEFR profiles was also found in the FGC lemmatized file since lemmas are, among others, singular forms of common nouns, not plural but the FGC only had occurrences of its plural form *centilitres*. In addition, FGC hyphenated words were considered as single types since no attempt was made to address the issue.

Synthesis of Results

This chapter discussed the findings both quantitative and qualitative of the study. The findings were presented in seven sections. The first four sections illustrated the overlap of the three resources. The remaining three sections highlighted types unique to each resource.

Initially, the CEFR profiles counted 41,255 unique types. After analysis and evaluation, acceptable A1 types constituted three-fourths (78.2%) of the initial A1 pool, whereas acceptable A2 and B1 types only about two-thirds more or less, 68.5 and 62.9% respectively. The greatest drop was seen with acceptable B2 types which were less than half of the initial B2 pool, i.e. 48.7%. In the end, only three out of five CEFR types were eligible for inclusion based on usage, i.e. 24,539 CEFR types or 59.5% of all CEFR types. The recommendation for inclusion totaled 43,635 types, an additional 2,380 types from the initial CEFR profile count. Nevertheless, 16,714 CEFR types (the remaining 40.5% of all CEFR types) were not included in this recommendation.

Findings from sections 2, 5, and 6 highlighted the types included in the FDF and FGC. Based on the data, 19,096 types were recommended for addition to the CEFR profiles (see Table 41 and 42 for the breakdown).

Table 41

Recommendations for Inclusion and Addition in the CEFR Profiles

Type Section	Criterion	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
1. Common to CEFR, FDF, and FGC	**a**	766	389	854	672	2,681
	b	39	26	102	121	288
	c	147	114	196	191	648
	d	2,907	2,129	4,062	3,934	13,032
2. Common Only to FDF and FGC *	**a**					1,570
	b					436
	c					9,181
3. Common Only to CEFR and FDF	**a**	600	773	1,838	2,452	5,663
4. Common Only to CEFR and FGC	**a**	10	18	19	56	103
	b	107	135	419	1,465	2,126
5. Unique to FDF *	**a**					6,930
6. Unique to FGC *	**a**					233
	b					746
7. Unique to CEFR	**a**	0	0	0	0	0
Acceptable Types Total		4,576	3,584	7,490	8,891	43,637
Percentage		78.2%	68.5%	62.9%	48.7%	

Comprehensive type lists for each study section can be consulted at the electronic address http://humanities.byu.edu/frnvocab (see Appendix C for filenames and their respective contents). There, readers can view the actual type content of revised CEFR French profiles. They encompass: (a) all acceptable types of each section by proficiency level, if already predetermined (15 files), (b) all questionable types of each section by proficiency level, if already predetermined (15 files), and (c) the final comprehensive list of inclusion and addition recommendations (1 file). These revisited and overhauled CEFR vocabulary profiles were

broken down by proficiency level A1, A2, B1, B2, and inclusive of acceptable types from sections 2, 5 and 6, not yet broken down by proficiency levels.

Table 42

Recommendations for Exclusion from the CEFR Profiles

Type Section	Criterion	A1 Types	A2 Types	B1 Types	B2 Types	Total Types
7. Unique to CEFR	**a**	126	354	1,585	4,668	6,733
6. Unique to FGC *	**a**					1,345
	b					301,222
5. Unique to FDF *	**a**					6,227
4. Common Only to	**a**	0	0	3	4	7
CEFR and FGC	**b**	138	253	808	2,382	3,581
3. Common Only to CEFR and FDF	**a**	1,013	1,043	2,027	2,310	6,393
2. Common Only to	**a**					0
FDF and FGC *	**b**					11
	c					451
1. Common to CEFR, FDF, and FGC		0	0	0	0	0
Questionable Types						
Total		1,277	1,650	4,423	9,364	325,970
Percentage		21.8%	31.5%	37.1%	51.3%	

* Note: FDF and FGC are not divided by proficiency levels so cells only contain totals.

Chapter 5: Conclusions

The study of French CEFR profiles compared with FDF and FGC yielded considerable quantities of information regarding lexical content. A substantial part of the resources overlapped as described in the analysis. The importance of this information is summarized in this chapter along with a short discussion of limitations and future work.

Summary of Findings and Recommendations

The empirical substantiation of French CEFR profile content was accomplished by comparing and evaluating CEFR types against two expansive and important present-day corpora, the FDF and the FGC. Task- and notion-based CEFR profiles have now been tested against frequency of usage data. The use of types as a unit of analysis gave a more precise idea of the quantity of vocabulary units needed to reach advanced language proficiency.

The findings of this corpus- and usage-based evaluation showed that current CEFR profiles were not totally representative of contemporary French. They allowed the elaboration of recommendations for honing CEFR content, and the establishment of curricular priorities. They also raised morphological awareness. Inclusion, addition, and exclusion criteria used here established an order of priority for learning and instruction.

The initial CEFR profile word count suggested that the CEFR contained twice as many wordforms as the FDF 5,000 most frequent lemmas. However, after checking for internal consistency and counting unique types in the CEFR resource, the French profiles had 23% less types than the FDF (41,255 compared to 53,511 types). Counting unique types, instead of traditional lemmas or word families, had the CEFR total jump six-fold from 6,407 units to 41,255 types, making the complexity of learning and teaching French more apparent. Of the total

CEFR types, 16.3% (6,733) were unique to the profiles and, thus, not frequently used. From the initial 41,255 total CEFR types, only 59.5% were retained (see Table 43, column 3). This evaluation also showed that progressive and cumulative vocabulary acquisition was not reflected in the present CEFR vocabulary profiles (see Table 43, column 2).

Among the non-CEFR types, a large number of types unique to FGC were observed but were not recommended for the new overhauled profiles. However, 96% of types common only to FDF and FGC were recommended for addition to the fine-tuned CEFR content (see Table 44).

Table 43

Summary of Recommendations for Inclusion to or Exclusion from the CEFR Profiles

CEFR Proficiency Level	Current CEFR Types	Types Recommended for Inclusion *	Types Recommended for Exclusion **
A1	5,853	4,576 (78.2%)	1,277 (21.8%)
A2	5,234	3,584 (68.5%)	1,650 (31.5%)
B1	11,913	7,490 (62,9%)	4,423 (37.1%)
B2	18,255	8,891 (48.7%)	9,364 (51.3%)
All 4 Levels	41,255	24,541 (59.5%)	16,714 (40.5%)

* Substantiated from 3 overlap sections: common core, CEFR & FDF, and CEFR & FGC
** Types unique to CEFR, or from CEFR & FDF or CEFR & FGC overlap

Table 44

Summary of Recommendations for Addition in the CEFR Profiles

Non-CEFR Types	Section Total	Types Recommended for Addition	Types Not Recommended for CEFR
2. FDF & FGC Overlap	11,649	11,187 (96.0%)	462 (4.0%)
5. Unique to FDF	13,157	6,930 (52.7%)	6,227 (47.3%)
6. Unique to FGC	11,913	979 (0.3%)	302,567 (99.7%)
Total Types	328,352	19,096 (5.8%)	309,256 (94.2%)

Thus the adjusted CEFR profiles would contain an additional 2,382 types compared to the initial profiles. They would, however, only include 59.5% of the original CEFR set, and add 19,096 new types from the FDF and FGC overlap (see Table 45).

Table 45

Recommendations for Newly Adjusted CEFR Profiles

Old CEFR Type Total	CEFR Type Inclusion	FDF and FGC Type Addition	New CEFR Type Total
41,255	24,541	19,096	43,637

This evaluation also demonstrated that current CEFR content would not suffice to prepare learners for advanced French. Francis and Kučera (1982) informed readers about vocabulary size and coverage, and found that learning 5,000 lemmas allowed 88.7% lexical coverage for English. The findings of this study reveal that the 5,000 most frequent French lemmas and all their derived inflections represented 78.6% of French text in a 23 million word corpus, and yielded 53,511 types, a ten-fold increase. By inference, the CEFR 41,255 French types would represent about 3,856 lemmas and cover 61% of French text. This finding shows that the CEFR profiles are indeed minimal vocabulary lists (as per profile developers). The coverage French profiles provide is insufficient for reading comprehension as demonstrated by Hu and Nation (2000). The number of types from proficiency levels A1 to B2 is inadequate to reach advanced proficiency. Ninety-five percent coverage allows for fluent reading and performance of advanced-level language tasks. Raising CEFR coverage to fluent reading level would require introduction of at least twice the amount of lemmas, i.e. 10,000 lemmas which would, in turn, require the learning of 100,000 French types or more. To reach this goal, between 25,000 to 30,000 types should be learned at each proficiency level. The current CEFR profiles introducing 41,255 types would only reach level A2 and not attain the B2-C1 intermediate to advanced threshold.

This study not only gave insight into French vocabulary quantity but also into vocabulary priority. One finding was that FDF and FGC types highlighted the preferences for usage of certain inflections over others (i.e., masculine over feminine, singular over plural, infinitive and

participle over other conjugated forms). The findings also drew attention to common and frequent, yet more domain-specific types, especially when reaching the intermediate proficiency levels, B1 and B2.

From FDF/FGC overlap or unique types, notions such as international relations including politics and military matters not found in the CEFR emerged. More specialized scientific and slang terminology transitioning from less to more frequent also surfaced. Types unique to the CEFR profiles drew attention to lexical dynamics, the movement of newer or antiquated forms from more to less frequent and vice-versa. This research supports Decoo's (2011) observations, i.e

> insufficient coordination efforts across instruction levels, redundancy, gaps in knowledge (. . .) e.g., no concretization of curriculum articulation to the vocabulary content level, (. . .) no mention of number and kinds of words learners are expected to know by the end of the instruction. (pp. 168-171)

Thus, questions raised vary from (a) How many types are students actually learning at beginning and intermediate level? (b) How are they counted? (c) How many types do teachers of French actually teach at each level? to (d) How long does it take for an average learner to learn 41,255 types? and (e) How many and which French types would allow a 95% coverage of French text?

The general recommendation would be to reorganize the French vocabulary profiles in such a manner that acceptable types of the sequenced sections of this study be taught and learned in the order they are listed, i.e. first the common core, then types common only to FDF and FGC, common only to CEFR and FDF, common only to CEFR and FDF, unique to FDF, and finally unique to FGC. They should be sequenced, and apportioned based on vocabulary size

determined by prior research and the number of types needed for 95% coverage in order to reach the B2-C1 proficiency threshold.

Other suggestions entail the following measures. First, increase the number of most frequent types learned at proficiency levels A1 and A2, compared to B1 and B2. This could be done by teaching common core types first until the determined number of types to be taught has been reached, then moving to the next best pool of frequent types, and so on. Second, lengthen instruction time spent at levels A1 and A2. This study shows that there are many types to learn. Learning these types will take time. Performance and evaluation, not time spent learning, will determine proficiency level. Third, increase overall and even out the number of types taught at each beginning and intermediate level so as to cumulatively reach the 95% French coverage, allowing fluent reading. Goal-centered instruction and careful planning focusing on the most common French types are more likely to lead to success and actual language acquisition. Fourth, account and test for number of types learned at each proficiency level from level A1 on. This usage-based strategy would keep language teachers and learners on target from the very start.

These findings are a clear contribution to the fields of instructional psychology and second language acquisition. They inform CEFR profile developers and authors, the European Language Policy Division, and French vocabulary test developers, in addition to learners, teachers, and administrators during formal instruction, language testing, and materials development and implementation. Applying these recommendations would lead to the restructuring of instructional content of French as a foreign language based on usage and improve the standardization of language instruction. Most frequent French words would be taught, recognized, and learned first to speed up language acquisition.

Limitations

The limitations of findings summarized above come from a choice made at the onset to look at lexical form rather than meaning. A morpho-semantic study of types would have multiplied the complexity already intrinsic to the present study of forms observed in the CEFR profiles, the FDF, and FGC. For instance, one single form might have had several meanings in one resource, and several different meanings in the other. This would have rendered the accounting of meanings and types quite difficult to carry out with substantial lexical resources. This study of most common French types does not consider specialized needs for specific contexts such as technological or occupational language requirements.

In the CEFR profiles, stakeholders have at their disposal a classification of words by general and specific notions. CEFR profile lists seem to give all words equal usefulness. Now, stakeholders have quantified usage-based information regarding these words, but they have to reconstruct the semantic categories these words go under, and possibly add new ones.

In addition, the FDF, despite its sizable balanced corpus, only satisfies the minimum requirement for reading comprehension. Its source corpus would need to be steadily enlarged, similar to the progressive development of a monitor corpus. It would require representative and balanced selections of general French, written and spoken, from all over the French-speaking world. The FGC, although close to a billion words strong, only represents one genre. As a consequence we find more specialized and advanced forms with its less frequent types. However, both the FDF and FGC were among the best accessible French lexical resources available at the time this study began.

Future Work

The study findings open the way to many new research avenues. Only a few are mentioned here.

First, the reverse engineering of types to lemmas would be in order. This manipulation would allow actual correspondance of frequency and ranking information for the FDF and FGC resources. FDF type frequency information could also be obtained. After these steps are completed, study types could be regrouped by frequency under each lemma and show their granularity.

Second, an order of CEFR notional priority based on usage could be determined. This kind of order would reinforce the implementation of recommendations mentioned above by presenting most frequent words common to the three study resources first, and so on.

Third, a semantic study could be undertaken of the FDF and FGC types recommended for addition in this evaluation, i.e. those not included in the overhauled CEFR which belong to sections 2, 5, and 6. This study would explore the semantic aspect of proposed forms, and clarify to which CEFR notions (general or specific) these FDF and FGC recommended types belong. It would explain whether these recommended additions fall into new semantic categories, and determine at what proficiency level they could be taught.

Fourth, the creation of a valid electronic word base using the study's recommended lexical content and permanently resubstantiated by a consequent monitor corpus could be undertaken. This word base would be endowed with frequency rankings and rank order. It would explicitly identify types by rank order at least up to the 95% coverage mark. Once recommended lexical content is presented via electronic interface, a range of operations and outcomes would become possible. Stakeholders of French language, including learners, teachers, evaluators, and

publishers, would focus on the task at hand: first learn the most frequent French types in order to understand most French texts and, then, concentrate on learning less frequent specialized words of their academic discipline or field of work. They would come to master language skills by

a. identifying more easily vocabulary included in (or excluded from) most frequent French words;

b. drawing from pedagogical materials à la carte and supplementing those skewed in one lexical domain or lacking in another.

This electronic tool would

a. provide a wide sample of texts with relevant words to be taught and learned;

b. draw attention to French lexical collocations and word order;

c. access rapidly varieties of French written and spoken outside of France;

d. check and assess lexical content of pedagogical materials against the pre-established lists;

e. benchmark the fine-tuned CEFR profiles and help assess lexical representativeness of pedagogical materials used for French L2 worldwide;

f. facilitate the redesign of French L2 curricula for beginner, intermediate, and advanced levels to better suit learners' and teachers' needs; and

g. enable the standardization of vocabulary learning and teaching at beginning and intermediate levels.

Fifth, an overlap comparison of substantiated French with substantiated language 1 forms (English or any other language; involving translation correspondences) could be conducted. Findings here could inform the design of an electronic bilingual or multilingual pedagogical corpus of balanced and representative texts, and increase the visibility of very frequent L1 and

L2 types. Stakeholders would here also be in a position to follow a word chronologically, and in concordance, and to find its meanings and its English or other language equivalents. It is envisioned that this corpus could be manipulated with a user-friendly query interface similar to COCA (Corpus of Contemporary American English) (Davies, 2008), and could also be used for evaluation purposes.

In an optimal design, the corpora mentioned above would function with added pedagogical features displaying more context or full-text files, and sound file options for intonation and pronunciation feedback. They could contain word difficulty indexes, text scanning and word analysis options to rapidly analyze lexical content, and immediately focus on most frequent word units, if unknown. They would facilitate planned teaching and learning of a mix of hard and easier frequent lexical items early on in the curriculum. They could change beginning and intermediate French courses into another part of French for specific purposes (FSP), the purpose here being to move learners from beginning to advanced level in as short a time as possible using content-based instruction. The subject matter content would be the identified most frequent words of the language corresponding to the 95% mark coverage. This type of usage-based French for basic foundational purposes could then be followed by the French module for specific purpose 1 (FSP1) and/or for purpose 2 (FSP2) and so on. These specific purposes would be determined by learners, employers, or expert teachers in their specific fields of work. Programs resulting from these exchanges should foster closer collaboration between linguistic experts in academic and professional fields.

A new definition for the French *Threshold level,* i.e. the transitioning from intermediate to advanced language proficiency, is proposed. It would correspond to the achievement of quantified partial to complete word knowledge of 95% of the most frequent French types. In

order to act according to this definition, and the findings and recommendations presented above, the Council of Europe, countries worldwide, educators and evaluators alike, need to encourage continued research that will build theory and link proficiency levels to quantitative and qualitative content specifications. Once provided with additional tested empirical evidence, they will be in a better position to make decisions regarding what words should be taught at what level.

References

English Vocabulary Profile - About the English Vocabulary Profile. (2011, November 19). Retrieved November 19, 2011, from English Profile - Setting professional standards for English language learning worldwide: http://vocabularypreview.englishprofile.org/staticfiles/about.html

Abdul-Rauf, S., & Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics* (pp. 16-23). Athens, Greece: Association for Computational Linguistics.

Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies.* New York: Routledge.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425-438.

Adolphs, S., & Schmitt, N. (2004). Vocabulary coverage according to spoken discourse context. In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 39-49). Amsterdam: Benjamins.

Alderson, C. (2007). Computer-adaptive language testing. In S. Granger, *Kaleidoscope - Optimizing the role of language in technology-enhanced learning* (pp. 1-3). Louvain: Université Catholique de Louvain (Belgium).

Alderson, J., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, S. (2004). *The development of specifications for item development and classification within the Common*

European Framework of Reference for Languages : Learning, teaching, assessment. Reading and listening. Final report of the Dutch DEF Construct Project. Available on request from J.C. Alderson, <c.alderson@lancaster.ac.uk>.

Allwright, L. R. (1984). Why don't learners learn what teachers teach?: The interaction hypothesis. In D. M. Singleton, & D. (. Little, *Language Learning in Formal and Informal Contexts.* (pp. 3-18). Dublin: Irish Association for Applied Linguistics.

Anderson, A., & Lynch, T. (1987). *Listening.* Oxford: Oxford University Press.

Bailey, K., & Nunan, D. (. (1996). *Voices from the language classroom.* Cambridge: Cambridge University Press.

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

Bauer, L., & Nation, I. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Beacco, J.-C. (1981). A demi-mot. *Le Français dans le Monde*, 74-75.

Beacco, J.-C. (2006). *Niveau A1.1 pour le français – Référentiel et certification (DILF) pour les premiers acquis en français.* Paris: Didier.

Beacco, J.-C., & Porquier, R. (2007). *Niveau A1 pour le français - Un référentiel.* Paris: Éditions Didier.

Beacco, J.-C., Blin, B., Houles, E., Lepage, S., & Riba, P. (2011). *Niveau B1 pour le français (apprenant / utilisateur indépendant) Niveau seuil.* France: Éditions Didier.

Beacco, J.-C., Lepage, S., Porquier, R., & Riba, P. (2008). *Niveau A2 pour le français - Un référentiel.* Paris: Éditions Didier.

Beacco, J.-C., Porquier, R., & Bouquet, S. (2004). *Niveau B2 pour le français - Un référentiel.* Paris: Éditions Didier.

Beauchemin, N., Margel, P., & Théoret, M. (1992). *Dictionnaire de fréquence des mots du français parlé au Québec: Fréquence, dispersion, usage, écart réduit.* New York: P. Lang.

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *SSLA*, 199-208.

Bogaards, P. (1994). *Le vocabulaire dans l'apprentissage des langues étrangères.* Paris: Éditions Didier.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition*, 321-343.

Breen, M. P. (1987). Contemporary paradigms in syllabus design, part I. *Language Teaching*, 81-91.

Brindley, G. (1987). Factors affecting task difficulty. In D. Nunan, & D. Nunan (Ed.), *Guidelines for the Development of Curriculum Resources.* Adelaide: National Curriculum Resource Centre.

Brisbois, J. E. (1995). Connections between first- and second-language reading. *Journal of Reading Behavior*, 565-584.

Brown, C. (1993). Factors affecting the acquisition of vocabulary: Frequency and saliency of words. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 263-286). Norwood, NJ: Ablex.

Brumen, M., Cagran, B., & Rixon, S. (2009). Comparative assessment of young learners' foreign language competence. *Educational Studies*, 269-295.

Brumfit, C. (1981). Notional syllabuses revisited: a response. *Applied Linguistics*, 90-92.

Brumfit, C. J., & Johnson, K. (. (1979). *The communicative approach to language teaching.* Oxford University Press.

Brunet, É. (1981). *Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la langue française.* Paris: Champion.

Brunet, É. (1987). L'exploitation des données du TFL. *Le Bulletin de l'EPI, 47*, 159-168.

Buxbaum, M. O. (2001). *1001 most useful French words.* Mineola, New York: Dover Publications.

California State Department of Education. (1986). *Statement on competencies in languages other than English expected of entering freshmen. Phase I--French, German, Spanish.* Sacramento: California State Department of Education, Bureau of Publications.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics, 1*(1), 1-47.

Candlin, C. N. (1987). Towards task-based learning. In C. N. Candlin, & D. Murphy (Eds.), *Lancaster Practical Papers in English Language Education* (Vol. 7, pp. 5-22). Englewood Cliffs, New Jersey: Prentice Hall.

Candlin, C., & Nunan, D. (1987). *Revised syllabus specifications for the Omani School English language curriculum.* Muscat: Ministry of Education and Youth.

Coady, J., & Huckin, T. N. (1997). *Second language vocabulary acquisition: A rationale for pedagogy.* New York: Cambridge University Press.

Cohen, A. (1986). Forgetting foreign language vocabulary. In B. Weltens, & K. de Bot, *Language attrition in progress.* Dordrecht: Foris.

Cohen, A., Glasman, B., Rosenbaum-Cohen, P., Ferrara, J., & Fine, J. (1979). Reading English for specialized purposes: Discourse analysis and the use of student informants. *TESOL Quarterly*, 551-564.

Cook, V. (1985). Language functions, social factors, and second language learning and teaching. *International Review of Applied Linguistics*, 177-198.

Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M., & Papo, E. (1976). *Un niveau-seuil.* Paris: Hatier.

Council of Europe. (1971). Linguistic content, means of evaluation and their interaction in the teaching and learning of modern language in adult education. *Symposium at Rüschlikon, Council of Europe Paper CCC/EES* (pp. 1-135). Strasbourg: Council of Europe.

Council of Europe. (2001). *Common European frame of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P. Arnaud, & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 1-12). London: MacMillan.

Coxhead, A. J. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Crombie, W. (1985). *Relational syllabuses.* Oxford: Oxford University Press.

Cummin, J. (1981). Interdependence threshold hypotheses.

David, A. (2008, December). Vocabulary breadth in French L2 learners. *Language Learning Journal, 36*(2), 167-180.

Davies, M. (2008). *Corpus of Contemporary American English*. Retrieved from COCA: http://corpus.byu.edu/coca/

Davies, M., & Face, T. L. (2006). Vocabulary coverage in Spanish textbooks: How representative is it? In N. Sagarra, & A. J. Toribio (Eds.), *Selected Proceedings of the 9th Hispanic Linguistics Symposium* (pp. 132-143). Somerville, MA: Cascadilla Proceedings Project.

De Jong, J. H. (2004). Comparing the psycholinguistic and the communicative paradigm of language proficiency. *International Workshop "Psycholinguistic and psychometric aspects of language assessment in the Common European Framework of Reference for Languages".* Amsterdam: University of Amsterdam.

De Saint Leger, D. (2009). Self-Assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, 158-178.

Decoo, W. (2011). *Systematization in foreign language teaching: Monitoring content progression.* London: Routledge.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: Johns Hopkins University Press.

ESOL Examinations. (2011). *Using the CEFR: Principles of good practice.* Cambridge: University of Cambridge.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Fulcher, G. (2008). Testing times ahead. *Liaison Magazine - Subject Centre for Languages, Linguistics and Area Studies, 1*, 20-24.

Furness, N. A. (1975). Not just a matter of luck: Some reflections on success and failure among first year students. *Modern Languages in Scotland*, 31-37.

Gardner, D. (. (2013). Technology and usage-based teaching applications. In C. A. Chapelle, *The Encyclopedia of Applied Linguistics.* Oxford, UK: Wiley-Blackwell.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in Children's narrative and expository reading materials. *Applied Linguistics*, 1-37.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.

Gass, S. (1989). Second language vocabulary acquisition. *Annual Review of Applied Linguistics, 9*, 92-106.

Gougenheim, G. (1958). *Dictionnaire fondamental de la langue française.* Paris: Librairie Marcel Didier.

Gougenheim, G., Michea, G. R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré) : Étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris: Didier.

Goulden, R., Nation, I. S., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*, 341-363.

Grabe, W. M. (1986). The transition from theory to practice in teaching reading. In F. Dubin, D. E. Eskey, & W. Grabe, *Teaching second language reading for academic purposes* (pp. 25-48). Reading, MA: Addison Wesley.

Graff, D. (2006). French Gigaword first edition. Linguistic Data Consortium: Philadelphia.

Graff, D., Mendonça, Â., & DiPersio, D. (2011). French Gigaword third edition. Linguistic Data Consortium, Philadelphia.

Harlow, L. L., & Muyskens, J. A. (1994). Priorities for intermediate-level language instruction. *The Modern Language Journal*, 141-154.

Hazenberg, S. (1994). *Een keur van woorden [A pick of words].* Amsterdam: Free University.

Hazenberg, S., & Hulstijn, J. H. (1992). Woorden op zicht: Woordselectie ten behoeve van het NT2-onderwijs. *Levende Talen, 467*, 2-7.

Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics, 17*, 145-163.

Henmon, V. A. (1924). *A French word book based on a count of 400,000 running words.* Madison, WI: University of Wisconsin.

Hirsch, D., & Nation, I. (1993). What vocabulary size is needed to read unsimplified text for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

Hu, M., & Nation, I. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). A diagnostic language assessment system for adult learners. In J. Alderson, *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies* (pp. 130-146). Strasbourg: Council of Europe.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, volume 91.

Imbs, P. (1971). *Dictionnaire des fréquences, vocabulaire littéraire des XIXe et XXe siècles.* Paris: Didier, Klincksieck.

Imbs, P. (1971-1994). *Trésor de la langue française.* Paris: CNRS, Gallimard.

Jacubowicz, C. (1989). Maturation or invariance of universal grammar principles in language acquisition. *Probus*, 283-340.

Juilland, A., Brodin, D., & Davidovitch, C. (1970). *Frequency dictionary of French words.* La Haye, Paris: Mouton.

Kellerman, E. (1991). Compensatory strategies in second language research: A critique, a revision, and some (non-)implications for the classroom. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood-Smith, & M. Swain (Eds.), *Foreign Second Language Pedagogy Research* (pp. 142-160). Clevedon, Avon: Multilingual Matters.

Kelly, P., Li, X., Vanparys, J., & Zimmer, C. (1996). A comparison of the perceptions and practices of Chinese and French-speaking Belgian university students in the learning of English: The prelude to an improved programme of lexical expansion. *ITL, Review of Applied Linguistics*, 275-303.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London and New York: Longman.

Kibby, M. W. (1977). Note on relationship of word difficulty and word frequency. *Psychological Reports, 41*, 12-14.

Kiss, G. (1968). Words associations and networks. *Journal of Verbal Learning and Verbal Behaviour, 7*, 707-713.

Kumaravadivelu, V. B. (1993). The name of the task and the task of naming: Methodological aspects of task-based pedagogy. In G. Crookes, & S. M. Gass, *Tasks in a pedagogical context* (pp. 69-96). Cleveland, UK: Multilingual Matters.

Kusseling, F., & Decoo, W. (2009). Europe and language learning: The challenges of comparable assessment. *34th European Studies Conference - University of Nebraska - Omaha.* Omaha: University of Nebraska.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Cleveland, UK: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J.-L. Arnaud, & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London: Macmillan Academic and Professional Ltd.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19*(2), 255-271.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-329.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Lazare, L. (1992). *French learner's dictionary.* New York: Living Language.

Legutke, M., & Thomas, H. (1991). *Process and experience in the language classroom.* Harlow: Longman.

Lété, B., Peereman, R., & Fayol, M. (2008). Consistency and word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study. *Journal of Memory and Language*, 952-977.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice.* Hove, England: Language Teaching Publications.

Lindqvist, C. (2010). Inter- and intralingual lexical influences in advanced learners' French L3 oral production. *IRAL*, 131-157.

Long, M. H., & Crookes, G. (1993). Units of analysis in syllabus design: The case for the task. In G. G. Crookes, & S. M. Gass, *Tasks in a pedagogical context* (pp. 9-44.). Cleveland, UK: Multilingual Matters.

Long, M. H., & Richards, J. C. (2007). Series Editors' Preface. In H. Daller, J. Milton, & J. Treffers-Daller, *Modelling and assessing vocabulary knowledge* (pp. xii-xiii). Cambridge: Cambridge University Press.

Lonsdale, D. (2012). *Personal comments.* Provo, Utah.

Lonsdale, D., & LeBras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners.* Abingdon, New York: Routledge.

Maun, I. (2009). Scaffolds for Reading in French: Lessons from History, Guidance for the Future? *Language Awareness*, 198-214.

McNichol, D. (1972). *A primer of signal detection.* London: Allen & Unwin.

Meara, P. (1990). Some Notes on The Eurocentres Vocabulary Size Tests. In J. Tommola, *Foreign language comprehension and production.* Turku: AFinLA.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P., & Jones, G. (1988). Vocabulary Size as a Placement Indicator. In P. Grunwell, *Applied Linguistics in Society.* London: CILT.

Meara, P., & Jones, G. (1990). *The Eurocentres vocabulary size tests.* Zurich: Eurocentres.

Mendonça, Â., Graff, D., & DiPersio, D. (2009). French Gigaword second edition. Linguistic Data Consortium: Philadelphia.

Meyer, C. F. (2002). *English corpus linguistics: An introduction.* Cambridge: Cambridge University Press.

Milton, J. (2006). French as a foreign language and the Common European Framework of Reference for Languages. *Crossing frontiers: Languages and the international dimension conference* (pp. 1-6). Cardiff University: CILT, the National Centre for Languages, the Subject Centre for Languages, Linguistics and Area Studies.

Milton, J. (2010). The development of vocabulary breacth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin, & I. Vedder, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211-231). eurosla.org: Eurosla.

Mondria, J.-A. (2006). Mythen over vocabulaireverwerving. *Levende Talen Tijdschrift, 7*(4), 3-11.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82.

Nation, P. (1990). *Teaching and learning vocabulary.* Boston: Heinle and Heinle.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Newton, J. (2001). Options for vocabulary learning through communication tasks. *English Language Teaching Journal*, 30-37.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 217-262.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Nunan, D. (1993). Task-based syllabus design: Selecting, grading and sequencing tasks. In G. Crookes, & S. M. Gass, *Tasks in a pedagogical context* (pp. 55-66). Cleveland, UK: Multilingual Matters.

Paulston, C. (1981). Notional syllabuses revisited: Some comments. *Applied Linguistics*, 93-95.

Platzer, H. (2006). English Competence among First-Semester Economics Students: An Empirical Investigation. *AAA, Arbeiten aus Anglistik und Amerikanistik*, 209-236.

Prabhu, N. S. (1987). *Second language pedagogy.* Oxford: Oxford University Press.

Qian, D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*(1), 1-19.

Read, J. (1989). *Towards a deeper assessment of vocabulary knowledge.* ERIC.ED 654 321.

Read, J. (1993). The Development of a New Measure of L2 Vocabulary Knowledge. *Language Testing, 10*(3), 355-371.

Richards, B., Malvern, D., & Graham, S. (2008). Word frequency and trends in the development of French vocabulary in lower-intermediate students during year 12 in English schools. *Language Learning Journal*, 199-213.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77-89.

Riley, P. (1982). Topics in communicative methodology: Including a Preliminary and Selective Bibliography on the Communicative Approach. In CRAPEL, *Mélanges Pédagogiques* (pp. 93-122). Nancy, France: Centre de Recherches et d'Applications Pédagogiques en Langues.

Robinson, P., Ting, C.-C., & Urwin, J.-J. (1996). Investigating second language task complexity. *RELC Journal, 26*, 62-79.

Rolland, J., & Picoche, J. (2008). Propositions pour un apprentissage systématique du lexique français au niveau A1 et au-delà. In J.-C. Beacco, *Niveau A1 et niveau A2 pour le français. Textes et références* (pp. 43-277). Paris: Didier.

Schinnerer-Erben, J. (1981). Sequencing redefined. *Practical Papers in English Language Education, 4*, 1-29.

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A study of the oral vocabulary of adults.* Brisbane: University of Queensland Press.

Schonenberg, N. (1988). *Étude comparative et didactique d'Un Niveau-seuil et du Français Fondamental. 1988 Master's thesis.* Antwerp: University Institution of Antwerp.

Schwenk, H., Fouet, J.-B., & Senellart, J. (2008). First steps towards a general purpose French/English statistical machine translation system. In P. o. Workshop, *Third ACL workshop on statistical machine translation* (pp. 119-122). Columbus, Ohio: Ohio State University.

Seedhouse, P. (1999). Task-based interaction. *English Language Teaching Journal, 53*(3), 149-156.

Shaw, T. (1997). Foreign language syllabus development: some recent approaches. *Language Teaching and Linguistics Abstracts*, 10-14.

Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher, 19*(2), 58-59.

Singleton, D. (1998). Age and the Second Language Lexicon. *Studia Anglica Posnaniensia*, 365-376.

Skehan, P. (1992). Strategies in second language acquisition. *Thames Valley University Working Papers in English Language Teaching, 1*. Thames Valley University.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics*, 38-62.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (2003). Task-based Instruction. *Language Teaching*, 1-14.

Stern, H. (1983). *Fundamental concepts of language teaching.* Oxford: Oxford University Press.

The British National Corpus. (n.d.). Retrieved from www.natcorp.ox.ac.uk/

Trim, J. (2011, August 4). *Welcoming address.* Retrieved December 10, 2011, from 2011 ACTFL CEFR Conference Report: http://www.actfl.org/files/PressReleasesOther/2011-ACTFL-CEFR.pdf

Upjohn, J. (1999). Exit Proficiency: The Proof of the Pudding. *ASp: La Revue du GERAS*, 305-322.

Van der Vliet, H. (1997). *Dingen onder woorden: conceptuele semantiek voor een computer-lexicon [Things into words: Conceptual semantics for a computer lexicon]*. Amsterdam: University of Amsterdam, IFOTT.

Van Ek, J. A. (1975). *Systems development in adult language learning: The threshold level in a European unit credit system for modern language learning by adults.* Strasbourg: Council of Europe.

Van Ek, J. A. (1976). *The threshold level for modern language learning in schools.* Strasbourg: Council of Europe.

Van Ek, J. A., & Trim, J. L. (1984). *Across the threshold.* Oxford: Pergamon Press.

Van Ek, J. A., Alexander, L. G., & Fitzpatrick, M. A. (1977). *Waystage: an intermediary objective below threshold level in a European unit/credit system of modern language learning by adults.* Strasbourg: The Council of Europe.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards, & B. Laufer, *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 175-189). Philadelphia: Johns Benjamins.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 281–300.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review, 53*, 13-40.

White, R. V. (1988). *The ELT curriculum, design, innovation and management.* Oxford: Basil Blackwell.

Widdowson, H. (1978). Notional-functional syllabuses: 1978, part 4. In C. H. Blatchford, & J. Schacter, *On TESOL '78* (pp. 32-35). Washington, DC: TESOL.

Widdowson, H. G. (1968). The teaching of English through science. In J. Dakin, B. Tuffen, & H. G. Widdowson, *Language in education: The problem in Commonwealth Africa and the Indo-Pakistan sub-continent* (pp. 115-175). London: Oxford University Press.

Widdowson, H. G. (1979). *Explorations in applied linguistics.* Oxford: Oxford University Press.

Wilkins, D. A. (1976). *Notional syllabuses - A taxonomy and its relevance to foreign language curriculum development.* London: Oxford University Press.

Williams, M., & Burden, R. L. (1997). *Psychology for language teachers: A social constructivist approach.* Cambridge: Cambridge University Press.

Yule, G., Powers, M., & Macdonald, D. (1992). The variable effects of some task-based learning procedures on L2 communicative effectiveness. *Language Learning*, 249-277.

Zavasnik, M. (2009). Higher education language teachers' perceptions of their competences. *Strani Jexici*, 91-103.

Zechmeister, E. B., D'Anna, C. A., Hall, J. W., Paus, C., & Smith, J. A. (1993). Metacognitive and other knowledge about the mental lexicon: Do we know how many words we know? *Applied Linguistics, 14*(2), 188-206.

Zipf, G. K. (1935). *The psycho-biology of language.* Boston: Houghton Mifflin Co.

Appendix A

Articles using the Gigaword Corpora

Bilingual word spectral clustering for statistical machine translation (Zhao et al, 2005); *Language models and reranking for machine translation* (Olteanu et al, 2006); *Parallel creation of Gigaword corpora for medium density languages – an interim report* (Halacsy, 2008); *First steps towards a general purpose French/English Statistical Machine Translation System* (Schwenk, Fouet, & Senellart, 2008); *Intelligent selection of language model training data* (Moore & Lewis, 2010); *Word lattices for morphological reduction and chunk-based reordering* (Hardmeier et al, 2010); *Transcriber driving strategies for transcription aid system* (Senay et al, 2010); *CMU syntax-based machine translation at WMT 2011* (Hanneman & Lavie, 2011), and *Hierarchical phrase-based Machine Translation at the Charles University for the WMT 2011 Shared Task* (Zeman, 2011).

Appendix B

Excerpt of the French Gigaword Corpus

```
<DOC id="AFP_FRE_19990201.0006" type="story" >
<HEADLINE>
Le Pakistan gagne en Inde: la victoire du sport
</HEADLINE>
<DATELINE>
MADRAS (Inde), 1er fév
</DATELINE>
<TEXT>
<P>
Le Pakistan a battu l'Inde dans le premier
test-match de cricket entre les deux frères ennemis sur le sol indien en 12
ans, une rencontre longtemps menacée par des extrémistes hindous, placée sous
très haute sécurité mais qui a vu la victoire du sport.
</P>
<P>
"Nous avons gagné. L'Inde aurait pu tout aussi bien gagner. Mais le vrai
vainqueur est le cricket lui-même", a déclaré le capitaine pakistanais Wasim
Akram après que le Pakistan eut battu l'Inde dimanche à Madras (sud-est) à
l'issue d'un test-match haletant de quatre jours.
</P>
<P>
L'équipe pakistanaise a effectué un tour d'honneur, saluée par une ovation
de quelque 40.000 spectateurs indiens.
</P>
<P>
"Chapeau bas au gens de Madras qui sont venus voir ce match, pour leur
comportement et leur discipline. Ils ont été absolument superbes, amicaux et
compréhensifs", a déclaré l'entraineur pakistanais, Javed Miandad.
</P>
<P>
"Quel test!", s'est exclamé lundi le journal Hindustan Times. "Le Pakistan,
mais aussi le cricket a triomphé".
</P>
<P>
Le Pakistan est en tournée de près de deux mois en Inde, la première depuis
1987. Un second test est prévu à New Delhi du 4 au 8 février.
</P>
<P>
Des militants d'un parti extrémiste hindou, le Shiv Sena, opposé à la venue
de l'équipe pakistanaise avaient endommagé un stade de New Delhi, saccagé le
siège de la fédération indienne de cricket et menacé de s'en prendre
physiquement aux joueurs pakistanais.
</P>
<P>
Sous la pression du gouvernement nationaliste hindou, conscient de
l'impopularité des menaces contre le cricket-roi en Inde, ce parti avait
renoncé à saboter la tournée juste avant l'arrivée de l'équipe pakistanaise le
21 janvier.
</P>
<P>
Le Shiv Sena accuse le Pakistan d'activités terroristes au Cachemire indien
où une guérilla musulmane soutenue selon New Delhi par le Pakistan a fait plus
de 24.000 morts depuis dix ans.
</P>
</TEXT>
</DOC>
```

Appendix C

Files listed on the website http://humanities.byu.edu/frnvocab

File Count	Filename	Type Section	Inclusion /Addition or Exclusion	Total Types
1	S1CoreA1TypIncl	1. Common to CEFR, FDF, and FGC	Inclusion	3,859
2	S1CoreA2TypIncl	1. Common to CEFR, FDF, and FGC	Inclusion	2,658
3	S1CoreB1TypIncl	1. Common to CEFR, FDF, and FGC	Inclusion	5,214
4	S1CoreB2TypIncl	1. Common to CEFR, FDF, and FGC	Inclusion	4,918
5	S2ComFDFFGCAdd	2. Common only to FDF and FGC	Addition	11,187
6	S3ComCEFRFDFA1TypIncl	3. Common only to CEFR and FDF	Inclusion	600
7	S3ComCEFRFDFA2TypIncl	3. Common only to CEFR and FDF	Inclusion	773
8	S3ComCEFRFDFB1TypIncl	3. Common only to CEFR and FDF	Inclusion	1,838
9	S3ComCEFRFDFB2TypIncl	3. Common only to CEFR and FDF	Inclusion	2,452
10	S4ComCEFRFGCA1TypIncl	4. Common only to CEFR and FGC	Inclusion	117
11	S4ComCEFRFGCA2TypIncl	4. Common only to CEFR and FGC	Inclusion	153
12	S4ComCEFRFGCB1TypIncl	4. Common only to CEFR and FGC	Inclusion	438
13	S4ComCEFRFGCB2TypIncl	4. Common only to CEFR and FGC	Inclusion	1,521
14	S5UniqFDFTypAdd	5. Unique to FDF	Addition	6,930
15	S6UniqFGCTypAdd	6. Unique to FGC	Addition	979
16	S7UniqCEFRA1TypExcl	7. Unique to CEFR	Exclusion	126
17	S7UniqCEFRA2TypExcl	7. Unique to CEFR	Exclusion	354
18	S7UniqCEFRB1TypExcl	7. Unique to CEFR	Exclusion	1,585
19	S7UniqCEFRB2TypExcl	7. Unique to CEFR	Exclusion	4,668
20	S6UniqFGCTypExcl	6. Unique to FGC	Exclusion	302,567
21	S5UniqFDFTypExcl	5. Unique to FDF	Exclusion	6,227
22	S4ComCEFRFGCA1TypExcl	4. Common only to CEFR and FGC	Exclusion	138
23	S4ComCEFRFGCA2TypExcl	4. Common only to CEFR and FGC	Exclusion	253
24	S4ComCEFRFGCB1TypExcl	4. Common only to CEFR and FGC	Exclusion	811
25	S4ComCEFRFGCB2TypExcl	4. Common only to CEFR and FGC	Exclusion	2,386
26	S3ComCEFRFDFA1TypExcl	3. Common only to CEFR and FDF	Exclusion	1,013
27	S3ComCEFRFDFA2TypExcl	3. Common only to CEFR and FDF	Exclusion	1,043
28	S3ComCEFRFDFB1TypExcl	3. Common only to CEFR and FDF	Exclusion	2,027
29	S3ComCEFRFDFB2TypExcl	3. Common only to CEFR and FDF	Exclusion	2,310
30	S2ComFDFFGCExcl	2. Common only to FDF and FGC	Exclusion	462
31	TypRecomIncl	from Sections 1, 2, 3, 4, 5, and 6	Inclusion/ Addition	43,637

Appendix D

Least Discrepant Types Not Found in CEFR Vocabulary Profiles (Increasing Order)

Types	FDF Rank	FGC Rank	Discrepancy |FDF-FGC|
dont	74	74	0
prôner	3,859	3,858	1
leur	35	32	3
soutenir	578	575	3
elle	38	43	5
lui	64	69	5
contester	1,974	1,981	7
transférer	2,128	2,135	7
dépêcher	3,771	3,762	9
juridiction	4,443	4,454	11
nous	31	45	14
nombreux	366	351	15
proposer	338	323	15
on	29	46	17
jusque	134	153	19
lutte	759	738	21
préjudice	4,858	4,880	22
soviétique	1,674	1,697	23
texte	631	607	24
infliger	2,730	2,705	25
je	22	49	27
fuir	1,960	1,931	29
imputer	4,397	4,427	30
apparemment	1,734	1,765	31
partenaire	1,077	1,111	34
identifier	1,426	1,461	35
incarner	3,574	3,539	35
déployer	1,718	1,680	38
complice	3,500	3,460	40
rendez-vous	1,873	1,833	40
destruction	1,921	1,880	41
inédit	4,590	4,631	41
armement	3,403	3,361	42
liaison	1,968	2,010	42
financement	1,671	1,628	43
affaire	170	125	45

(Table Continued)

Types	FDF Rank	FGC Rank	Discrepancy |FDF-FGC|
finance	1,677	1,632	45
baser	1,712	1,666	46
engagement	1,042	1,088	46
général	147	100	47
sinistre	3,578	3,531	47
celui	45	95	50
homosexuel	3,501	3,551	50
reconstruction	3,111	3,059	52
prédécesseur	3,948	3,895	53
vacance	1,726	1,782	56
détruire	928	985	57
humour	3,950	4,010	60
tenter	347	287	60
regrouper	2,477	2,540	63
agir	211	275	64
démocratique	1,380	1,315	65
nazi	3,053	3,118	65
restructuration	3,331	3,398	67
coupable	1,442	1,511	69
prévenir	1,207	1,135	72
territoire	698	624	74
parmi	389	464	75
crime	819	897	78
outre	974	896	78
quasiment	3,538	3,617	79
procédure	993	1,076	83
enfuir	3,804	3,892	88
comédie	3,373	3,284	89
prélèvement	3,662	3,753	91
convoquer	2,520	2,428	92
souhaiter	403	311	92
téléphonique	2,356	2,264	92
fiscal	1,637	1,730	93
duc	4,804	4,899	95

Appendix E

Most Discrepant Types Not Found in CEFR Vocabulary Profiles (Decreasing Order)

Types	FDF rank	FGC rank	Discrepancy |FDF-FGC|
chier	4245	133844	129599
hé	3871	121114	117243
présentement	4691	95190	90499
spécification	4642	89575	84933
ouais	1928	85806	83878
rayer	4631	72936	68305
dix-neuvième	3997	71507	67510
euh	889	67932	67043
chéri	2880	63283	60403
emmerder	4676	52368	47692
vérificateur	4257	45595	41338
hein	2076	43157	41081
ha	4965	45189	40224
putain	2704	41884	39180
interface	3240	40566	37326
premièrement	3587	39821	36234
merde	2376	36745	34369
lunette	4207	36592	32385
salaud	4869	32646	27777
là-dedans	3796	31023	27227
ah	1405	27699	26294
bordel	4529	29737	25208
distorsion	4231	28397	24166
compression	4995	29054	24059
connerie	4402	27937	23535
nominal	3482	26539	23057
exprès	4999	27807	22808
deuxièmement	3730	26206	22476
effrayer	3828	26274	22446
pis	3579	24914	21335
relire	4170	24922	20752
camoufler	4704	25282	20578
trame	3457	23448	19991
rigoler	4994	24594	19600
infiniment	4550	23891	19341
interaction	4970	23535	18565

(Table continued)

Types	FDF rank	FGC rank	Discrepancy |FDF-FGC|
murmurer	4730	23286	18556
récepteur	4552	22415	17863
écu	3760	21576	17816
crête	4895	22441	17546
affreux	4369	21796	17427
soixante-dix	4887	22295	17408
autochtone	2387	19448	17061
écrier	4865	21575	16710
compliment	4922	21572	16650
terrien	4835	21172	16337
quoique	3243	19566	16323
inhérent	4672	20939	16267
blâmer	4803	21036	16233
mec	2358	18030	15672
spécifier	3086	18753	15667
dilemme	4798	20346	15548
primitif	3345	18777	15432
bosser	4532	19830	15298
gosse	3631	18908	15277
insensible	4901	20141	15240
indépendamment	4296	19204	14908
vulgaire	4886	19702	14816
réputé	3768	18582	14814
concurrentiel	4179	18794	14615
rente	4876	19478	14602
con	2817	17399	14582
ensuivre	4942	19395	14453
assurément	4535	18896	14361
léguer	4058	18212	14154
sou	4051	17938	13887
au-dessous	3965	17661	13696
soumission	4578	18224	13646
tâcher	4043	17460	13417
ironique	4500	17869	13369
advenir	4469	17755	13286
civiliser	4227	17422	13195
par-dessus	3800	16923	13123
ignorance	4055	17114	13059
balayer	3687	16737	13050

(Table continued)

Types	FDF rank	FGC rank	Discrepancy |FDF-FGC|
néant	3707	16574	12867
humble	4680	17353	12673
énoncer	3541	15944	12403
omettre	4177	16575	12398
inévitablement	4726	17012	12286
surplus	3104	15358	12254
majesté	4174	16422	12248
repenser	3834	16073	12239
soi-disant	3637	15807	12170
vôtre	4065	16197	12132
englober	4188	16302	12114
compréhensible	4774	16771	11997
superficiel	4740	16597	11857
insensé	4614	16439	11825
foutre	1890	13663	11773
instinct	4482	16119	11637
rigide	4741	16367	11626
convenable	4410	16025	11615
différencier	4620	16114	11494
avantageux	4391	15854	11463
sincèrement	3516	14953	11437
théologie	4468	15867	11399
idiot	3556	14733	11177
honorable	893	12068	11175
paradoxe	4279	15419	11140
tantôt	3013	14076	11063
eh	1692	12752	11060
carrément	4009	14906	10897
comptabilité	3837	14695	10858
vingt-quatre	4149	14982	10833
leadership	4000	14817	10817
cynique	3547	14328	10781
préjugé	3679	14418	10739
pertinent	3348	14074	10726
notoire	4781	15483	10702
aucunement	4408	15054	10646
incomber	4617	15210	10593
obséder	4663	15198	10535
blague	4822	15326	10504

(Table continued)

Types	FDF rank	FGC rank	Discrepancy |FDF-FGC|
guetter	4852	15349	10497
empresser	4927	15345	10418
décent	4628	15031	10403
explicite	3798	14185	10387
contrevenant	4118	14427	10309
balancer	4548	14850	10302
insérer	3970	14229	10259
attentivement	3962	14196	10234
caresser	4861	15074	10213
fardeau	3954	14128	10174
planification	4710	14875	10165
truc	1991	12083	10092
tisser	4320	14378	10058
extraction	4262	14295	10033
énumérer	4848	14868	10020
déformer	4386	14399	10013
affectation	4868	14880	10012
productivité	2901	12907	10006