A Frequency Dictionary of French

A Frequency Dictionary of French is an invaluable tool for all learners of French, providing a list of the 5000 most frequently used words in the language.

Based on a 23-million-word corpus of French which includes written and spoken material both from France and overseas, this dictionary provides the user with detailed information for each of the 5000 entries, including English equivalents, a sample sentence, its English translation, usage statistics, and an indication of register variation.

Users can access the top 5000 words either through the main frequency listing or through an alphabetical index. Throughout the frequency listing there are thematically organized lists of the top words from a variety of key topics such as sports, weather, clothing, and family terms.

An engaging and highly useful resource, the Frequency Dictionary of French will enable students of all levels to get the most out of their study of French vocabulary.

Deryle Lonsdale is Associate Professor in the Linguistics and English Language Department at Brigham Young University (Provo, Utah). Yvon Le Bras is Associate Professor of French and Department Chair of the French and Italian Department at Brigham Young University (Provo, Utah).

Routledge Frequency Dictionaries

Other books in the series:
A Frequency Dictionary of Mandarin Chinese
A Frequency Dictionary of German
A Frequency Dictionary of Portuguese
A Frequency Dictionary of Spanish
A Frequency Dictionary of Arabic (forthcoming)

A Frequency Dictionary of French

Core vocabulary for learners

Deryle Lonsdale and Yvon Le Bras

LONDON AND NEW YORK

First published 2009 by Routledge 2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada by Routledge 270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2008.

To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

British Library Cataloguing in Publication DataA catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication DataLonsdale, Deryle. A frequency dictionary of French : core vocabulary for learners / Deryle Lonsdale, Yvon Le Bras. p. cm. Includes index. 1. French language—Word frequency—Dictionaries. I. Lonsdale, Deryle. II. Title. PC2691.L66 2009 443′.21—dc19 2008042400

ISBN 0-203-88304-7 Master e-book ISBN
ISBN10:0-415-77531-0 (pbk)
ISBN10:0-415-77530-2 (hbk)
ISBN10:0-203-88304-7 (ebk)
ISBN13:978-0-415-77531-1 (pbk)
ISBN13:978-0-415-77530-4 (hbk)
ISBN13:978-0-203-88304-4 (ebk)

Contents

Thematic vocabulary list vi
Series preface vii
Acknowledgments ix
Abbreviations x
Introduction 1
References 8
Frequency index 9
Alphabetical index 204
Part of speech index 258

Thematic vocabulary lists

1	Animals	9
2	Body	16
3	Food	23
4	Clothing	30
5	Transportation	37
6	Family	44
7	Materials	51
8	Time	58
9	Sports	65
10	Natural features and plants	72
11	Weather	79
12	Professions	86
13	Creating nouns – 1	93
14	Relationships	100
15	Nouns – differences across registers	107
16	Colors	114
17	Opposites	121
18	Nationalities	128
19	Creating nouns – 2	135
20	Emotions	142
21	Adjectives – differences across registers	149
22	Verbs of movement	156
23	Verbs of communication	163
24	Use of the pronoun “se”	170
25	Verbs – differences across registers	178
26	Adverbs – differences across registers	186
27	Word length	195

Series preface

There is a growing consensus that frequency information has a role to play in language learning. Data derived from corpora allows the frequency of individual words and phrases in a language to be determined. That information may then be incorporated into language learning. In this series, the frequency of words in large corpora is presented to learners to allow them to use frequency as a guide in their learning. In providing such a resource, we are both bringing students closer to real language (as opposed to textbook language, which often distorts the frequencies of features in a language, see Ljung 1990) and providing the possibility for students to use frequency as a guide for vocabulary learning. In addition we are providing information on differences between frequencies in spoken and written language as well as, from time to time, frequencies specific to certain genres.

Why should one do this? Nation (1990) has shown that the 4,000–5,000 most frequent words account for up to 95 per cent of a written text and the 1,000 most frequent words account for 85 per cent of speech. While Nation's results were for English, they do at least present the possibility that, by allowing frequency to be a general guide to vocabulary learning, one task facing learners – to acquire a lexicon which will serve them well on most occasions most of the time – could be achieved quite easily. While frequency alone may never act as the sole guide for a learner, it is nonetheless a very good guide, and one which may produce rapid results. In short, it seems rational to prioritize learning the words one is likely to hear and use most often. That is the philosophy behind this series of dictionaries.

The information in these dictionaries is presented in a number of formats to allow users to access the data in different ways. So, for example, if you would prefer not to simply drill down through the word frequency list, but would rather focus on verbs, the part of speech index will allow you to focus on just the most frequent verbs. Given that verbs typically account for 20 per cent of all words in a language, this may be a good strategy. Also, a focus on function words may be equally rewarding – 60 per cent of speech in English is composed of a mere 50 function words.

We also hope that the series provides information of use to the language teacher. The idea that frequency information may have a role to play in syllabus design is not new (see, for example, Sinclair and Renouf 1988). However, to date it has been difficult for those teaching languages other than English to use frequency information in syllabus design because of a lack of data. While English has long been well provided with such data, there has been a relative paucity of such material for other languages. This series aims to provide such information so that the benefits of the use of frequency information in syllabus design can be explored for languages other than English.

We are not claiming, of course, that frequency information should be used slavishly. It would be a pity if teachers and students failed to notice important generalizations across the lexis presented in these dictionaries. So, for example, where one pronoun is more frequent than another, it would be problematic if a student felt they had learned all pronouns when

they had learned only the most frequent pronoun. Our response to such issues in this series is to provide indexes to the data from a number of perspectives. So, for example, a student working down the frequency list who encounters a pronoun can switch to the part of speech list to see what other pronouns there are in the dictionary and what their frequencies are. In short, by using the lists in combination a student or teacher should be able to focus on specific words and groups of words. Such a use of the data presented here is to be encouraged.

Tony McEnery and Paul Rayson Lancaster, 2005

References

Ljung, M. (1990)A Study of TEFL Vocabulary. Stockholm: Almqvist & Wiksell International.

Nation, I.S.P. (1990)Teaching and Learning Vocabulary. Boston: Heinle and Heinle.

Sinclair, J.M. and Renouf, A. (1988) “A Lexical Syllabus for Language Learning”. In R. Carter and M. McCarthy (eds) Vocabulary and Language Teaching London: Longman, pp. 140–158.

Page ix

Acknowledgments

We are first and foremost grateful to Mark Davies for proposing that we undertake this work, and for his occasional guidance and suggestions throughout its duration. This work also would not have been possible without the help of our able and hard-working student research assistants at Brigham Young University: Fritz Abélard, Amy Berglund, Katharine Chamberlin, and Ben Sparks.

The first author would like to thank his French instructors throughout his formative years, particularly France Levasseur-Ouimet and Gérard Guénette. He also acknowledges the inspiring influence of past colleagues in translation and lexicography including Greg Garner, Benoît Thouin, Brian Harris, Robert Good, Alain Danik, and Claude Bédard. He dedicates this book to his parents, to his wonderfully supportive wife Daniela, and to Walter H. Speidel whose own pioneering work in corpus-based computerized lexicography stands as an example for all of us who work in this field.

The second author wishes to thank Philippe Hamon, Bernard Quemada, and Réal Ouellet, his professors at the University of Rennes, the University of Paris III, and Laval University, who instilled in him the desire to study and teach the French language and literature. He dedicates this book to his parents and especially to his wife Hoa for her continued support and encouragement in his professional endeavors.

Page x

Abbreviations	Categories	Example
adj	adjective	1026 lourd adj *heavy*
adv	adverb	1071 certainement adv *certainly*
conj	conjunction	528 puisque conj *since*
det	determiner	214 votre det *your*
intj	interjection	889 euh intj *er, um, uh*
n	noun	802 absence nf *absence*
nadj	noun/adjective	4614 insensén adj *insane*
prep	preposition	389 parmi prep *among*
pro	pronoun	522 lui-même pro *himself*
v	verb	1014 confirmer v *to confirm*

Features on categories		Example
f	feminine	1011 armée nf *army*
i	invariable	1324 après-midi nmi *afternoon*
m	masculine	707 signe nm *sign*
pl	plural	3654 dépens nmpl *expense*
(f)	no distinct feminine	3770 apte adj(f) capable
(pl)	no distinct plural	3901 croix nf(pl) *cross*

Introduction

The value of a frequency dictionary for French

Today French is the second most taught and widespread second language globally, behind English. Yet, surprisingly, there is no current corpus-based frequency dictionary of the French language. The present dictionary is meant to address this shortcoming, and is part of a series that includes other highly useful dictionaries for Spanish (Davies, 2006) and Portuguese (Davies & Preto-Bay, 2008). As such it is similar in intent, approach, structure, and content to its predecessors. As noted below, some modifications have also been made to make it more usable for English speakers, who do constitute the largest group of speakers on the planet.

The purpose for this book is to prepare students of French for the words that they are most likely to encounter in the "real world". It is meant to help alleviate the phenomenon encountered all too often in dictionaries and language primers where word lists are introduced based on intuitive or unverifiable notions of which words might conceivably be most useful for students to acquire, and in which order. The dictionary is designed primarily as a reference work which could be used in concert with standard classroom curricular materials or used on an individual study basis. Ideas on how to carry out this integration have been noted in the previous dictionaries noted above.

Contents of the dictionary

This is first and foremost a frequency dictionary. The principal information concerns the 5,000 most frequent words in French as determined in the process described below. This information is arranged in three different formats: (i) a main frequency listing, which begins with the most frequent word (with associated information) followed by the next most frequent word, and so forth; (ii) an alphabetical index of these words, and (iii) a frequency listing of the words organized by part of speech, and (iv) thematic lists grouping some of the words into related semantic classes. Each of the entries in the main frequency listing contains the word itself, its part(s) of speech (e.g. noun, verb, adjective, etc.), a context reflecting its actual usage previously in French, an English translation of that context, and summary statistical information about the usage of that word. Some or all of this information is likely to be highly useful for language learners in different settings.

The vocabulary itself was derived from a corpus, or body, of French texts. The corpus we collected was assembled specifically for this work and totals millions of words, half of them reflecting transcriptions of spoken French and the other half written French texts. Since the dictionary is focused primarily on frequency and usage, the words do not have associated with them any pronunciation guides, etymological history, or domain-specific usage information. The dictionary is also focused on single words, which is a crucial but not exclusive consideration in language learning; to extensively address fixed word expressions such as collocations and idioms would be beyond the scope of this dictionary.

The dictionary, then, is designed as an instrument for helping students acquire a core vocabulary of French words in various ways, including based on their observed frequency in recent French language usage. The versatility in its organization should presumably allow its use in a wide range of language learning scenarios.

Previous frequency dictionaries for French

French dictionaries are plentiful and widely varied in content, so one might wonder whether another

dictionary is necessary. A short survey of existing dictionaries should suffice to illustrate why this one was developed.

Two landmark frequency dictionaries have been produced for French. One (Henmon 1924) was based on 400,000 words of text, and the other (Juilland et al. 1970) derives from a study of 500,000 words.

Information on the words contained in those lists, though, was minimal, and the ability to handle more sizable corpora has since – of course – been vastly improved with computer technology.

Other word reference lists have been developed largely for scholarly purposes and hence not very accessible to the average learner. Brunet (1981) focuses on development of French vocabulary over time based on the superb Trésor de la Langue Française (Imbs 1971-1994). Beauchemin et al. (1992) focus only on the French spoken in Quebec. All of these resources require some effort to use effectively.

Some lexical resources are at the disposal of French language learners through the Internet, such as the ARTFL FRANTEXT and TLFi resources. The subscription costs and on-line access methods are sometimes less practical than having a reasonably sized dictionary like this one at one's fingertips.

Finally, some helpful recent beginner dictionaries exist, though each has its own limitations. Recent ones by Oxford University Press (2006), Living Language (Lazare 1992), and Dover Publications (Buxbaum 2001) list from 1001 to 20,000 "most useful" words but give no rationale for how they were selected. Another venerable work by Gougenheim (1958) lists 3500 basic French words with related information including definitions, but which are entirely in French and hence challenging for the beginner.

Our dictionary seeks to combine the best from this tradition of French lexical research while at the same time avoiding these shortcomings. Its presentation design and the rationale and methodology for selecting the contents reflect what we believe to be the state of the art in corpus research, text processing, and lexicography.

The corpus and its annotation

Our dictionary is derived from a corpus of some 23,000,000 French words that have been assembled from a wide variety of sources. As mentioned above, half of this total reflects a collection of transcriptions from oral or spoken French, while the other half reflects French in its textual or written form. Reflecting a desire to make our dictionary a modern representation of the French language, we have included no materials that date before the year 1950.

We did not try to proportion our data based on geographical region or demographics, but we did try to achieve some balance across genres; however, this balance is not perfect. It is also important to note that some of our content from particular sources was exhaustive whereas in other cases it was selectively or randomly sampled; in other words, only parts of the material were used because there was too much content and hence the risk of skewing coverage of particular areas.

The spoken text portion of the corpus was made up of approximately 11.5 million words. These words were pulled from such various forms such as transcripts of governmental debates/hearings, telephone calls, and face-to-face dialogues. There were also transcripts of interviews with writers, entertainment figures, business leaders, athletes, academicians and other media personnel. And

finally we made use of movie scripts/subtitles and theatrical plays.

The written text portion of the corpus was also made up of roughly 11.5 million words. This part of the corpus was assembled from newswire stories, daily and weekly newspapers, newsletters, bulletins, business correspondence, and technical manuals. Magazines such as popular science and other technical publications were used. We also targeted different genres of literature such as fiction/nonfiction essays, memoirs, novels and more.

Table 1 gives a more detailed listing of the composition of the corpus.

Corpus standardization and annotation

Collection of the corpus involved much work in what has been called corpus standardization or text preprocessing. Given the wide range of sources for the corpus, they involved many different file types, character encodings, and formatting conventions. For example, the documents used a wide range of character representations and formats such as EBCDIC, MACROMAN, ISO, UTF-8, and HTML. In many cases unneeded material such as images, advertisements, or templatic information had to be stripped out, a process called document scrubbing.

Each type of transcription or text document was then processed so that the paragraphs, sentences, words, and characters were identified and encoded in a standard way to enable further processing, a process called tokenization. The scrubbing and tokenization processes involve linguistic issues that had to be addressed, such as deciding on how to break up

Table 1 Composition of 23 million word French corpus

Spoken	
Approx. # of words	Sources
175	Conversations (3)
3,750,000	Canadian Hansard (4)
3,020,000	Misc. interviews/transcripts (5)
1,000,000	European Union parliamentary debates (6)
855	Telephone conversations (7)
470	Theatre dialogue/monologue (8)
2,230,000	Film subtitles (9)

TOTAL
11,500,000

Written	
3,000,000	Newswire stories (10)
2,015,000	Newspaper stories (11)

4,734,000	Literature (fiction, non-fiction) (12)
434	Popular science magazine articles (13)
1,317,000	Newsletters, tech reports, user manuals (14)

TOTAL
11,500,000

GRAND TOTAL
23,000,000

3 The French portion of the C-ORAL-ROM corpus (Cresti & Moneglia 2005).

4 Aligned Hansards of the 36th Parliament of Canada; for more information consult http://www.isi.edu/natural-language/download/hansard/.

5 Miscellaneous transcripts of interviews with various business, political, artistic, and academic personalities mined from hundreds of Internet sites. Many were from media sites such as French television studios (e.g. www.tf1.fr and www.france2.fr), publishing houses (www.lonergan.fr), popular culture websites (e.g. www.evene.fr), and business information portals (e.g. http://www.journaldunet.com).

6 A small random sampling from the French portion of the Multilingual Corpora for Cooperation (MLCC) corpus. See resource W0023 at www.elda.fr.

7 Aligned transcribed training data from the ESTER Phase 2 evaluation campaign; downloaded from http://www.irisa.fr/metiss/guig/ester/.

8 A small random sampling of extracts from theatrical works posted at various sites including www.leproscenium.fr.

9 Selected portions of several film subtitles from Jörg Tiedemann's OPUS corpus; downloaded from http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php.

10 A tiny random sampling of stories from the French GigaWord corpus; for more information see http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T17.

11 A sampling from newspaper articles on the Internet from journalism sites throughout the French-speaking world (e.g. www.lemonde.fr, www.ledevoir.com).

12 Samples and complete short works of fiction and non-fiction works from various publishing houses (e.g. www.edition-grasset.fr, www.lonergan.fr) and Web virtual libraries (e.g. www.gutenberg.org).

13 A variety of articles from popular science magazine sites on the Internet (e.g. www.pourlascience.com, www.larecherche.com, etc.).

14 A variety of technical report and newsletter articles including weather bulletins, user manuals, business newsletters, and banking correspondence. Some of these materials are sampled from the

French portions of the European Corpus Initiative (see http://wwww.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T17).

words separated by hyphens (dis-moi vs. week-end) and apostrophes (l'homme vs. aujourd'hui). Some documents had accented upper letters whereas others did not, so the process of case folding – or reducing capitalized words to their lower-case form – was also nontrivial. Many special symbols including degree signs, ellipsis punctuation, currency symbols, bullets, and dots also required standardization. To perform all of this work we used several file conversion programs as well as our own Perl scripts, Unix tools (e.g. make, awk, grep, sort, uniq, join, comm), and SGML/HTML/XML parsers.

Once the corpus was standardized, it was then necessary for us to assign to each word its part of speech; in other words, whether it functions as a noun, a verb, an adjective, and so forth. Currently there are about a dozen different part of speech taggers for the French language, each with its own theoretical framework, implementation approach, and set of tag encodings to flag the relevant parts of speech for each word. In this work we installed and tested several of these taggers. In our case we found that each tagger had its own strengths and weaknesses and that by combining several of them and merging the results in a postprocessing stage we could create our own tagging procedure and tagset to produce the best results for our purpose. We also performed a certain amount of editing and correcting tagging results by hand for the most common tagging errors, though for the entire corpus a thorough examination of each word would have been prohibitively time-consuming and costly.

It was also necessary to perform a morphological analysis of each word in the corpus to find its base form, or lemma. For example, the second word in the sentence "Je suis heureux." is a verb conjugation of the verb "être", which is its base form or lemma. Similarly, pronouns with regular inflections (e.g. "il" to "ils"), adjectives, and determiners with variant forms were combined together. The lemmatization process was necessary for our frequency computations, to be described below. Various lemmatization programs exist for French, and in fact some of them perform both part of speech tagging and lemmatization at the same time. In this stage, too, there were challenges that we had to overcome. For example, many words are morphologically ambiguous, having several possible lemmas, such as the verb form "suis" having both "être" and "suivre" as possible lemmas, depending on the particular instance. Another difficulty is deciding when non-finite forms (i.e. past and present participles and infinitives) function more as verbs or as other parts of speech (especially nouns, adjectives). Again we found that combining some of the most popular programs and postprocessing the results ended up being the most helpful for our purposes.

Target vocabulary identification and description

With the whole corpus standardized and annotated, it was possible to compute word frequencies and identify the most-used words. Counting words in a corpus can be done in several ways. We have chosen to collapse all of the variant forms of the same word and sum them up together. For example, the word "pour" is a conjunction or preposition and occurs in two other forms across the corpus: "Pour" and "POUR". Summing up all occurrences of the variant forms of this word we arrive at a total count of 151,709. Similarly, plural forms of nouns are normally reduced to their singular form, verb conjugations are reduced to their infinitive form, and inflected adjectives are reduced to the masculine singular form, as is done in other French dictionaries. For example, throughout the corpus there are 25 different forms of the verb "déterminer" including inflections and variant forms such as "déterminerait", "détermine", "déterminons", and "Déterminez"; all of these were combined with their counts into the infinitive form.

Our target vocabulary list is thus formed from the top 5000 scoring lemmas in the corpus. In identifying these top 5000 lemmas, some items (such as proper nouns and punctuation) were rejected. However, one more refinement was necessary in identifying the top 5000 words. Experience in corpus linguistics has shown that the raw frequency count for all variants of a word turns out not to be the best measure of its usefulness. Consideration must be made of how widely a word is spread across the different parts of a corpus.

Exactly quantifying how widely a word is spread across a document or corpus has been a thorny problem in corpus linguistics. If a given word occurs very frequently in one part of the corpus (e.g. the spoken part) but not elsewhere, it might be desirable to discount that word's raw frequency so that it becomes a little less "important" in comparison to

other less-frequent words. Literally dozens of approaches have been taken over the last decades to come up with workable solutions. One of the most promising, and the one used in the compilation of this book, is called the "deviation of proportions", or DP (Gries, 2008).

The DP measure looks at the proportion of a term's occurrence across various "slices" of a corpus, taking into account the size of each slice. Each word's final calculation involves three steps: (i) summing up all of the occurrences of that word's for each slice and normalizing it against that word's overall frequency in the whole corpus, called the "observed proportion"; (ii) normalizing each corpus slice with respect to the size of the whole corpus, called the "expected proportion"; (iii) computing the absolute difference between observed and expected proportions, summing them up, and dividing by 2. The result is a measure between 0 and 1, where 0 means the word is distributed evenly across the corpus slices and 1 means it is restricted to narrow parts of the corpus.

While helpful in describing word distribution across a corpus, the DP measure is only one metric, and for the purposes of this dictionary it was necessary to combine it with the raw frequency. Thus we computed, for each lemma, its frequency divided by its DP. The result determined the ranking of each lemma and hence its final appearance and relative order in the top 5000 words in the vocabulary. For example, all forms of the word "avoir" sum up to a frequency of 405,020 and its DP score is 0.11533. Its ranking score is thus 405020/0.115363, or 3,510,831.029. This is the sixth highest score among all of the lemmas, so this word places sixth in the ranked list.

Finally, the DP values are somewhat unwieldy as long numbers behind a decimal point. To solve this problem we mapped these values to a much more intuitive set of integers ranging from about 27 to 100. These numbers are called dispersion codes. The mathematical calculation for obtaining a dispersion code from its corresponding DP measure involves an exponential function: 100*exp-DP. Values approaching 100 indicate that the word is quite evenly distributed across the corpus; values below 50 indicate words that are limited to only certain narrow portions of the corpus.

Though these computations are somewhat technical, the general intuition is that the words in this dictionary are ranked by the summed frequency of all of their variant forms, tempered by how well they are spread across various portions of the corpus.

Once the terms were identified, additional information had to be collected to construct the associated entries.

Developing associated information

Providing parts of speech was done through a combination of automatic and manual methods. The values were derived from (i) the part of speech tags provided from the lemmatization process described above; (ii) popular lexical databases for French lexical information (e.g. BDLEX1); and (iii) hand-editing of the merged and accumulated results.

Glossing the terms was a completely manual effort. An intuitive effort was made to give as much of the core meaning(s) as possible while at the same time avoiding the temptation to be exhaustive.

The next stage involved finding a suitable usage context for each word. In each case the usage context comes from the corpus itself, so that it represents an illustration of natural French, the way a French-speaking person would use the word. Equally important was the need to find contexts that were clear, short, self-contained, and indicative of the core meaning of the word. Ideally, the contexts should also contain as few words as possible that are not covered in the dictionary elsewhere. To find the contexts, a computer-generated list of possible contexts was prepared for each word, and then scored automatically according to these criteria. We then manually chose from among these lists the best context for each word.

Like glossing, generating English translations for the usage contexts was also a human effort. Each context was taken in isolation and, often using the English glosses that had been prepared, a translation was entered manually. Some texts already had English translations from previous work and hence could have been extracted manually using word-alignment techniques, but we purposely chose to not use these techniques so as to assure that the translations were “fresh” in each instance.

1 See http://www.irit.fr/PERSONNEL/SAMOVA/decalmes/ IHMPT/ress_ling.v1/rbdlex_en.php.

Finally, we compiled the thematic lists. In each case the content of the list was done using a combination of automatic and manual techniques. For semantic subject areas (e.g. food and weather terms) hierarchical lexical databases (e.g. French WordNet2) were used to locate the terms’ position in a taxonomy of semantic field areas. A parallel effort of hand-selecting relevant terms was also carried out, and the results were merged together.

All of these results have been combined into a comprehensive database (we used both mySQL and Microsoft Access) that enables versatile retrieval of relevant information.

In conclusion, this dictionary is calibrated to the learners’ needs, and organized in such a way that is easy for the reader. Corpus linguistics is at the core of the effort, but a wide array of human skills and computational linguistic techniques were vital in the process.

The main frequency index

The frequency index is the main portion of this dictionary: it contains a ranked list of the top 5000 lemmas in French, starting with the highest-scoring one and progressing to the lowest-scoring word. Each entry has the following information:

ranked score (1, 2, 3…), headword, part(s) of speech, English gloss, sample context, English translation of sample context, dispersion value, raw frequency total, indication of register variation

For example, here is the entry for the word “aimer”:

242 **aimer** v *to like, love*

* tu sais que je t’aime -- you know I love you
71 | 10085 −n

This entry shows that the word (and all of its related forms) ranks 242nd among all French words in terms of combined frequency and dispersion. The part-of-speech code shows that it’s a verb. Two possible English glosses are “to like” and “to love”. One context from the corpus is shown, which uses one of the related forms of this verb: “aime”. An English translation for the usage context then appears. Next, the number “71” flags the dispersion value for the word on a scale from 27 to 100; the word and its forms are reasonably evenly spread across the corpus. The number “10085” indicates the raw frequency, or how many times the word and its related forms occur in the corpus. Finally, a register code –n indicates that this word is noticeably infrequent in nonfiction.

Here are some additional notes for the items appearing in the entries.

The part(s) of speech

Several categories have been combined to increase readability. For example, nadj *signifies a word that can be either a noun or an adjective. Marking for major features is also provided, such as for gender (nm for masculine, nf for feminine), number (pl for non-distinct plurals), and invariable words that don’t inflect (e.g. adji). Some nouns have both genders. In this dictionary participles that have drifted semantically from their core meaning, or that have acquired a status that makes them more like adjectives or nouns, have been listed separately. Examples of such words include “reçu (receipt)”, “fabricant (manufacturer)”, and “âgé (old)”.*

The English gloss

The gloss is meant to be indicative only – it’s not a complete listing of all possibilities. This is not an exhaustive bilingual dictionary. Many of these words also participate in idioms, fixed expressions, collocations, or multi-word expressions. These meanings are not included in the glosses since the focus is on single words. The glosses are written in standard American English. In certain parts of the dictionary (e.g. in the thematic lists) only shortened forms of the glosses are used.

The French usage context

As noted above, all of the usage contexts come from the corpus itself. In selecting them the goal was to find contexts that illustrate clearly the core meaning of the word as concisely as possible. Contexts will sometimes unavoidably include words that are not in the top 5000 words, as well as occasional idiomatic usages. The contexts are taken verbatim, with only very infrequent correction (e.g. spelling errors). Capitalization is (for the most part) neutralized to improve readability. Sometimes the contexts are not always grammatically correct, especially when taken

2 See http://www.illc.uva.nl/EuroWordNet/.

from a spoken language transcript where speech errors, non-standard usage, and non-prescribed forms are common (e.g. “j’sais pas” vs. “je ne sais pas”). Finally, the contexts reflect real-world usage, and hence may not always be factually or politically correct. No editorial endorsement or philosophical conclusions should be ascribed to the authors on the basis of the contexts used.

The English translation of the context

Translations are rendered into American English, with it attendant spelling and vocabulary choices. An attempt was made to project the register, style, and structure. of the source context into its translation. However, translations were focused more on meaning than structure, so some translations are fairly loose and sometimes creative. On occasion a translation may involve a word whose part of speech is different than its related word appearing in the English gloss. Since the glosses are only meant to be indicative and not exhaustive, the translations may involve words not shown in the glosses, but usually the relationship is readily apparent. Sometimes idiomatic usages or collocations may be used in a translation to show how flexibly the word can be used. It is important to note that the contexts were translated in standalone fashion. Since the context was isolated from its surrounding material before translation, reference of pronouns, articles, etc. is not guaranteed to perfect match the meanings in the source documents. For example, a context “Je lui téléphone” might either be translated “I’m telephoning him” or “I’m telephoning her”. Clearly one translation would be most appropriate to the context when viewed in its original source file, but to supply such is not necessarily our intention here.

The statistical and register information

The last line of each entry has two numbers divided by a vertical bar. The first is the dispersion value discussed above. The second is the raw frequency count for all of the variants of the entry’s headword. Some words also have a register code that specifies the word’s distribution across registers. The three registers and their codes are spoken (±s), literature (±l), and non-fiction (±n). A positive value for some register means that the word occurs in the top 5% of the expected frequencies for the words in that register, when compared against the other two registers. Conversely, a negative value means that the word occurs in the bottom 5% of the expected frequencies for the words in that register. For example the first-person pronoun “je” has a −n register code, indicating that it occurs comparatively very infrequently in the nonfiction register. On the other hand, a very imagery-laden descriptive adverb like “brusquement” has the codes +l and −s, meaning that it is very infrequent in spoken language but very frequent in literature.

Thematic vocabulary (“call-out boxes”)

A number of thematically-grouped words are given in tables that are placed throughout the frequency index. These include lists of terms for such semantic classes as animals, body parts, foods, colors, nationalities, and professions. Other tables give data on grammatical questions (e.g. use of the pronoun “se”), word length, and variation of word usage across the three registers (spoken, literature, and nonfiction). When glosses appear in these lists, they may only represent a portion of the glosses given in the main frequency list. In addition, sometimes words are ambiguous in their meaning; for example, the word “poisson (fish)” is both an animal and a food. In cases where the word’s usage exhibits a clear preference for one sense over another, it will only appear in the list associated with the preferred sense. In other cases, though, where no clear preference exists, the word may appear in both relevant thematic lists.

Alphabetical and part of speech indexes

The alphabetical index gives an alphabetical listing of all of the words listed in the previous section. Each entry in this chapter includes: (1) the lemma (2) the part of speech (3) a basic English equivalent, and (4) the word’s score in this dictionary.

The part of speech index gives a listing of the words from the frequency index, this time arranged by “parts of speech”. Each category lists the lemmas by their score in decreasing frequency of

occurrence. The alphabetical index can be used to link a given word with its score.

Page 8

References

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

Brunet, É. 1981.Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la langue française. Paris: Champion. (Travaux de linguistique quantitative, 46).

Buxbaum, M.O. 2001.1001 Most Useful French Words. Mineola, NY: Dover Publications.

Davies, M. 2006.Frequency Dictionary of Spanish: Core Vocabulary for Learners. New York: Routledge.

Davies, M. and Preto-Bay, A.M.R. 2008.Frequency Dictionary of Portuguese: Core Vocabulary for Learners. New York: Routledge.

Galarneau, A. 2002. Les dictionnaires de langue française. Dictionnaires d'apprentissage. Dictionnaires spécialisés de la langue. Dictionnaires de spécialité. International Journal of Lexicography (15)3:246–248.

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

Gries, S.T.forthcoming. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics.

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

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

Juilland, A., Brodin, D., and Davidovitch, C.1970.Frequency Dictionary of French Words. La Haye, Paris: Mouton.

Lazare, L 1992.French Learner's Dictionary. New York: Living Language.

Page 9

Frequency index

rank frequency (501, 502…), headword,part of speech, English equivalent

• sample sentence

range count | raw frequency total, indication of major register variation

1 **le** det, pro *the; him, her, it, them*
    * vive la politique, vive l'amour -- long live politics, long live love

89 | 2359662

2 **de** det, prep *of, from, some, any*
* il ne rêve que d'argent et de plaisirs -- he only dreams of money and pleasure
88 | 1665907

3 **un** adj, det, nm, pro *a, an, one*
* je me suis cassé un ongle -- I broke one of my fingernails
95 | 421500

4 **à** prep *to, at, in*
* ils restent à l'école le plus longtemps possible -- they remain at school as long as possible
93 | 557546

5 **être** nm,v *to be; being*
* tout le monde veut être beau -- everybody wants to be beautiful
91 | 514562

6 **et** conj *and*
* et les larmes se remirent à couler -- and the tears started flowing again
93 | 364443

7 **en** adv,prep,pro *in, by*
* je suis retournée en Espagne en septembre -- I returned to Spain in September
94 | 242952

8 **avoir** nm,v *to have*
* on était six donc tu peux pas avoir une conversation -- there were six of us so you can't have a conversation
89 | 405020

9 **que** adv,conj,pro *that, which, who, whom*
* c'est un soldat. mais que fait-il ici? -- it's a soldier. but what's he doing here?
88 | 348428

10 **pour** prep *for, in order to*
* elle jouait pour gagner -- she played to win
93 | 151709

11 **dans** prep *in, into, from*
* je reviendrai dans dix minutes -- I will return in 10 minutes
93 | 161033

12 **ce** det,pro *this, that*
* je ne déteste pas cet homme -- I do not detest this man
87 | 307421

13 **il** pro *he, it*
* allez voir s'il est blessé -- go see if he is injured
86 | 251585

14 **qui** pro *who, whom*

* je ne sais pas à qui m'adresser -- I don't know who to talk to
89 | 160867

15 **ne** adv *not*
* nous ne faisons pas du très bon travail -- we are not doing very good work
86 | 195309

16 **sur** adj,prep *on, upon*
* t'avais une chance sur un million -- you had one chance in a million
92 | 97798

1 **Animals**
animal 1002 M *animal*
poisson 1616 M *fish*
chien 1744 M *dog*
cheval 2220 M *horse*
oiseau 2435 M *bird*
bête 2591 F *beast*
vache 2768 F *cow*
chat 3138 M *cat*
monstre 3353 M *monster*
virus 3382 M *virus*
bœuf 3914 M *ox*
loup 3927 M *wolf*
porc 4036 M *pig*
mouton 4175 M *sheep*
rat 4290 M *rat*
poule 4321 F *hen*
souris 4328 F *mouse*
singe 4739 M *monkey*
ours 4800 M *bear*
bétail 4842 M *livestock*
cochon 4947 M *pig*
canard 5295 M *duck*
lion 5413 M *lion*
serpent 5574 M *snake*
puce 5788 F *flea*
lapin 5833 M *rabbit*
papillon 5979 M *butterfly*
dragon 6054 M *dragon*
chèvre 6074 F *nanny goat*
saumon 6287 M *salmon*
moule 6520 F *mussel*

17 **se** pro *oneself, himself, herself, itself, themselves*
* avec ce traité, le Japon se rapproche des Etats-Unis -- with this treaty, Japan brings itself closer to the U.S.
88 | 144707

18 **pas** adv,nm(pl) *not, n't; footstep*

* non, ne touchez pas! -- no, don't touch it!
83 | 161746

19 **plus** adv *more, no more*
* il est considérablement plus jeune que moi -- he's considerably younger than I am
90 | 85987

20 **pouvoir** nm,v *can, to be able to*
* tu peux jouer de la guitare électrique -- you can play electric guitar
90 | 78074

21 **par** prep *by*
* il s'y trouvait par hasard -- he found himself there by accident
87 | 99628

22 **je** pro *I*
* je suis contente de vous revoir -- I am happy to see you again
73 | 227259-n

23 **avec** prep *with*
* vous voulez aller au ciné avec moi? -- do you want to go to a movie with me?
91 | 66056

24 **tout** adv,det,nadj,pro *all, very*
* comme vous voyez, tout est propre -- as you see, everything is clean
86 | 95071

25 **faire** nm,v *to do, make*
* qu'est-ce qu'il fait? -- what's he doing?
85 | 99587

26 **son** det,nm *his, her, its; sound; bran*
* un ami ingénieur du son m'aide pour les arrangements -- a sound engineer friend of mine helped me with the arrangements
81 | 116681

27 **mettre** v *to put, place*
* je peux me mettre à votre table? -- may I sit at your table?
96 | 19654

28 **autre** det,nadj(f),pro *other*
* il y a un autre problème -- there's another problem
91 | 40519

29 **on** pro *one, we*
* on tire et on pose les questions ensuite -- we shoot first and ask questions later
80 | 87982

30 **mais** adv,conj,intj *but*
* je ne suis pas riche, mais je connais la vérité -- I'm not rich, but I know the truth
83 | 69661

31 **nous** pro *we, us*
* nous devons nous défendre nous-mêmes -- we must defend ourselves
78 | 89100

32 **comme** adv,conj *like, as*
* Tony et moi, on est comme des frères -- Tony and I, we're like brothers
87 | 49608

33 **ou** conj *or*
* il en reste du café ou pas? -- is there some coffee left or not?
86 | 49714

34 **si** adv,conj,nmi *if, whether*
* aujourd'hui, notre économie va si mal -- today our economy is going so poorly
87 | 46439

35 **leur** det,adj(f),pro *them, their, theirs*
* l'énergie solaire assurait leur survie -- solar energy assured their survival
87 | 39904

36 **y** adv,nmi,pro *there*
* c'est certain qu'on va y aller -- it's for certain that we'll be going there
83 | 55889

37 **dire** nm,v *to say*
* je décrochais le téléphone sans rien dire à personne -- I picked up the telephone receiver without saying anything to anyone
77 | 66657-n

38 **elle** pro *she, her*
* j'étais fou amoureux. elle m'aimait bien -- I was head-over-heels in love. she loved me a lot
78 | 66136

39 **devoir** nm,v *to have to, owe; duty*
* je dois travailler sans la moindre entrave -- I must work without the least bit of hindrance 83 | 46491
40 avant adji,adv,nm,prep *before*

* tu vas te pencher en avant -- you're going to lean forward
94 | 14109

41 **deux** det,nmi *two*
* il prend le train deux fois par semaine pour affaires -- he takes the train on business twice per week
87 | 31727

42 **même** adj(f),adv,pro *same, even, self*
* ils ne s'excusent même pas -- they don't even excuse themselves
85 | 37784

43 **prendre** v *to take*
* elle lui prit la main -- she took him by the hand

90 | 24323

44 **aussi** adv,conj *too, also, as*
* je rêvais aussi de beaucoup voyager -- I also dreamed of traveling a lot
88 | 26219

45 **celui** pro *that, the one, he, him*
* tu es celui que je respecte le plus -- you're the one I respect the most
87 | 28015

46 **donner** v *to give*
* j'aurais donné ma vie pour lui -- I would have given my life for him
90 | 19581

Page 11
47 **bien** adj *i,adv,nm well*
* tout va bien maintenant -- everything's going well now
81 | 39477

48 **où** adv,pro *where*
* on ne dit pas où il vit -- they aren't saying where he lives
87 | 27126

49 **fois** nf(pl) *time, times*
* lève la main une fois -- raise your hand one time
92 | 15724

50 **vous** pro *you*
* ici, vous avez une personnalité publique -- here, you are a public personality
65 | 77127-n +s

51 **encore** adv *again, yet*
* tu as encore menti à ta femme -- you lied once again to your wife
89 | 19772

52 **nouveau** adj,nm *new*
* il a construit une nouvelle vie ici -- he made a new life for himself here
88 | 22005
53 **aller** nm,v *to go*
* tu devrais aller te coucher, tu as l'air vanné -- you should go to bed, you look wiped out
73 | 50452

54 **cela** pro *that, it*
* cela demande de l'intégrité et du courage -- that requires integrity and courage
73 | 50891-n +s

55 **entre** prep *between*
* je marche entre les maisons -- I'm walking between the houses
86 | 23028

56 **premier** det,nadj *first*
* est-elle la première épouse, la deuxième? -- is she the first wife, the second one?

84 | 27061

57 **vouloir** nm,v *to want*
* tu veux faire ton chemin. c'est bien -- you want to continue on your way. that's OK
79 | 36467

58 **déjà** adv *already*
* les rues étaient déjà pleines de monde -- the streets were already full of people
93 | 11298

59 **grand** adv,nadj *great, big, tall*
* tu es plus grand que je pensais -- you're taller than I thought
87 | 21583

60 **mon** det *my*
* t'aurais pu rencontrer mon copain -- you could've met my buddy
70 | 55084-n

61 **me** pro *me, to me, myself*
* reviens me voir dans cinq ou six ans -- come back to see me in five or six years
65 | 63357-n

62 **moins** adji(pl),adv,nm(pl),prep *less*
* il avait moins d'excuses encore que ses complices -- he had even less excuses than his accomplices did
90 | 14730

63 **aucun** det,adj,pro *none, either, neither, not any*
* trop d'argent et aucun goût -- too much money and no taste
93 | 9641

64 **lui** pro *him, her*
* mais j'ai confiance qu'en lui -- but I have confidence in him
75 | 38286

65 **temps** nm(pl) *time*
* sur le plan spirituel, le temps n'existe pas -- on the spiritual level, time does not exist
86 | 20420

66 **très** adv *very*
* j'ai été bonne? très, très bonne -- was I good? very, very good
82 | 26324

67 **savoir** nm,v *to know*
* je ne savais plus quoi dire -- I didn't know what to say any more
78 | 32739

68 **falloir** v *to take, require, need*
* il ne faut pas être raciste, point -- there's no need to be a racist, at all
84 | 22844

69 **voir** v *to see*