Behavior Research Methods, Instruments, & Computers
2004, 36 (1), 156-166

MANULEX: A grade-level lexical database from French elementary school readers

BERNARD LÉTÉ
INRP, CNRS (UMR 6057) and Université de Provence, Aix-en-Provence, France

LILIANE SPRENGER-CHAROLLES
CNRS (UMR 8606) and Université de Paris 5, Paris, France

and

PASCALE COLÉ
CNRS (UMR 5105) and Université de Savoie, Chambéry, France

This article presents MANULEX, a Web-accessible database that provides grade-level word frequency lists of nonlemmatized and lemmatized words (48,886 and 23,812 entries, respectively) computed from the 1.9 million words taken from 54 French elementary school readers. Word frequencies are provided for four levels: first grade (G1), second grade (G2), third to fifth grades (G3–5), and all grades (G1–5). The frequencies were computed following the methods described by Carroll, Davies, and Richman (1971) and Zeno, Ivenz, Millard, and Duvvuri (1995), with four statistics at each level (*F*, overall word frequency; *D*, index of dispersion across the selected readers; *U*, estimated frequency per million words; and *SFI*, standard frequency index). The database also provides the number of letters in the word and syntactic category information. MANULEX is intended to be a useful tool for studying language development through the selection of stimuli based on precise frequency norms. Researchers in artificial intelligence can also use it as a source of information on natural language processing to simulate written language acquisition in children. Finally, it may serve an educational purpose by providing basic vocabulary lists.

This article presents MANULEX,[1] the first French linguistic tool that provides grade-based frequency lists of the 1.9 million words found in first-grade, second-grade, and third- to fifth-grade French elementary school readers. The database contains 48,886 nonlemmatized entries and 23,812 lemmatized entries. It was compiled to supply the French counterpart to such works on the English language as Carroll, Davis, and Richman's (1971) *American Heritage Word Frequency Book* and Zeno, Ivenz, Millard, and Duvvuri's (1995) more recent *Educator's Word Frequency Guide*.

Corpus-based word frequency counts are established as robust predictors of word recognition performance. Consequently, they are widely used in psycholinguistic research. Burgess and Livesay (1998) found them in almost 20% of the papers published in the main experimental psychology reviews. The word frequency effect, first noted by Cattell (1886), is one of the earliest empirical observations in cognitive psychology. Cattell demonstrated that the frequency of occurrence of a word in a language affects even the most basic processing of that word (its speed of recognition). Since this pioneering work, word frequency has been a persisting subject of study for investigators concerned with word recognition: High-frequency words are recognized more quickly and with greater accuracy than are low-frequency words, whatever the measure and task considered (for a review on word frequency effects, see Monsell, 1991). In fact, all current models of word recognition must incorporate word frequency in their activation mechanisms (for a review, see Jacobs & Grainger, 1994). Since the 1980s, for example, word frequency counts have been used mostly in connectionist modeling to simulate language development (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). As has been described by Zevin and Seidenberg (2002), in these models, knowledge is encoded as weights on connections between units, which reflect the cumulative effects of exposure to all the words. Learning the meaning of a word is thought to be dependent on exposure to that word in its linguistic contexts, and corpus-based word frequency counts are interpreted as a reflection of such individual experiences with a word.

Thus, a crucial variable for understanding language development and, particularly, the reading process is the

Support for this research was provided by two national grants, Cognitive Sciences (COG-192) and Schools and Cognitive Sciences (2001, AL16b), and by subsidies from the National Institute of Pedagogical Research (INRP, 1997, 1998, 1999). The authors thank Patrick Bonin and two anonymous reviewers for their insightful comments on a previous version of the article. We are grateful to Michel Fayol for constructive discussions, to Sébastien Pacton for his help in the early stages of the project, and to Vivian Waltz for her assistance with English. Correspondence concerning this article should be addressed to B. Lété, Laboratoire Parole et Langage, Université de Provence, 29 Avenue Robert Schuman, 13621 Aix-en-Provence Cedex 1, France (e-mail: lete@inrp.fr).

nature of the written vocabulary that children experience. A grade-based quantification of written material directed at children is a valuable tool for psycholinguists who wish to access written language learning in children. The use of word frequency norms computed from adult corpora raises problems, because they reflect the final state of the lexical system of an individual, but not the dynamics of how that system is built. Zevin and Seidenberg (2002) recently pointed out this methodological problem. Studying the age-of-acquisition (AoA) effect, they found that Zeno et al.'s (1995) grade-based frequency counts were more closely correlated with word-reading latencies than were earlier counts, such as those obtained by Kučera and Francis (1967). They explained these results by the fact that the grade-based frequency norms in Zeno et al.'s database were computed from child-targeted texts, and they emphasized the need for precise frequency norms to gain access to child language development.

Before the presentation of MANULEX, the state of the art in English and French lexical databases will be briefly described below.[2]

BACKGROUND: LEXICAL DATABASES FOR ENGLISH AND FRENCH

The history of lexicographical studies based on quantitative data is not recent, one of the most often quoted ancestors being Käding (1897), who established a lexical database as an aid to the shorthand recording of political, administrative, and business-related speeches in German. It was also for pragmatic purposes, educational in this case, that Thorndike (1921) established his English teacher's word book. A few years later, Thorndike participated in a conference held in New York that focused on the establishment of a basic English for language teaching and language diffusion, the core idea being to determine a basic vocabulary, which required determining word frequencies (Thorndike, 1932). The main goal of these early studies was quite different from that of today's studies in the same field, the aim of which is mainly to create tools for linguistic and psycholinguistic research, the most frequently quoted tools for American English being the word frequency lists in the Thorndike–Lorge count (Thorndike & Lorge, 1944), the Brown corpus (Kučera & Francis, 1967), and the *American Heritage Word Frequency Book* (Carroll et al., 1971).

Short History of French Lexical Databases

In the French-speaking countries, word frequency tables began to be established in the early 20th century, mainly to help teachers. The first one was proposed by Henmon (1924), who wanted to scientifically determine which words were the most common words and establish their degree of frequency. This work was based mostly on texts selected from the French literature of the second half of the 19th century. Ten years later, Vander Beke (1935) studied a wider corpus by introducing some nonliterary works, particularly scientific texts and newspaper articles. The main merit of this work was that it took into account a cross-corpus word dispersion index in such a way that a word appearing once in five different corpora, for example, was considered more significant than a word appearing 10 times in only one corpus.

The above corpora were established mainly from texts for adults. One of the first works including mainly texts written for—and even by—children was presented in Aristizabal's (1938) doctoral dissertation based on 4,125 schoolchildren's written productions (and also 1,400 adult letters). The Dubois and Buyse (1940/1952) scale was derived from this work: 3,724 words from the Aristizabal corpus were dictated to 59,469 elementary school children and were classified into 43 steps on the basis of which words were correctly spelled. The scale was updated into 40 steps by Ters, Mayer, and Reichenbach (1969). In the same line, Dottrens and Massarenti (n.d.) in Switzerland conducted a study that was based on Prescott's (1929) work, and Préfontaine and Préfontaine (1968) in Québec first established a list derived from 5- to 8-year-olds' spoken language, which served as a basis for selecting words for their teaching-to-read method.

The idea of a basic French vocabulary based on spoken corpora was also behind the book by Gougenheim, Michéa, Rivenc, and Sauvageot (1964), which gives the frequencies of 7,995 everyday conversation words, established from 275 recorded conversations (only the 1,063 most frequent words were retained for the publication). Catach, Jejcic, and the HESO group (1984) drew from this work and from two others based on written texts—Imbs (1971) and Juilland, Brodin, and Davidovitch (1970), the originality of the latter being that it takes into account the frequency of lemmatized and nonlemmatized words—to establish a list of the most frequent French words and their most frequent flexional forms (2,357 entries).

This short presentation shows that French researchers in child language development and French school teachers have few tools with which to do their jobs. These "databases" are very outdated, but they are still in use because no other alternative exists. In addition, these linguistic materials were extracted from children's written productions or adults' speech. As was pointed out by Smolensky (1996), the fact that children's linguistic ability in production lags dramatically behind their ability in comprehension poses a long-standing conceptual dilemma for studies of language acquisition. Children's productions do not reflect their competence in the same way as that assumed for adults, and there is a much greater competence/performance gap for children. As a result, the use of the Dubois–Buyse scale or the Catach lists to select items for studying word recognition in French, for example, raises several methodological and theoretical issues. However, these early works paved the way for French adult language computerized databases, which are presented below.

Current Computerized Language Corpora and Lexical Databases

English language. In English, computerized lexical databases have been available since the early sixties. The Brown corpus of standard American English was the first of the modern, computer-readable, general corpora. It was compiled by Kučera and Francis (1967) at Brown University and was intended to be a standard reference. The corpus consists of 1 million words from American English texts printed in 1961 and sampled from 15 different text categories.

The British national corpus (BNC) is a 100-million word collection of samples of written (90%) and spoken (10%) language from a large range of sources and is designed to represent a wide cross-section of current British English. The BNC is a unique snapshot of the English language and is designed to render possible almost any kind of computer-based research on language. Leech, Rayson, and Wilson (2001) recently published a word frequency book derived from the BNC that includes frequencies for writing and for present-day speech.

Some corpora, such as the MRC psycholinguistic database (Coltheart, 1981), have been compiled in specific lexical databases. The MRC contains 150,837 English words likely to be used in psycholinguistic research and provides information about 26 different linguistic properties. It was established from the following different sources in order to take into account most of the factors influencing lexical processing (for a review of some of these effects on word recognition, see Taft, 1991): the associative thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973), Jones's (1963) pronouncing dictionary of the English language, Paivio's ratings of the concreteness, imagery, and meaningfulness of words (Paivio, Yuille, & Madigan, 1968), Gilhooly and Logie's (1980) ratings based on AoA, imagery, concreteness, familiarity, and ambiguity measures, the Colorado norms, which deal with word meaningfulness (Toglia & Battig, 1978), the word frequency counts of Kučera and Francis (1967) and Thorndike and Lorge (1944), and the *Shorter Oxford English Dictionary* database (Dolby, Resnikoff, & MacMurray, 1963).

The *American Heritage Word-Frequency Book* (Carroll et al., 1971) is based on a survey of U.S. schools. It contains 5.09 million words from publications widely read by American schoolchildren 7–15 years of age. The set of 86,741 distinct words was created from 500-word samples taken from over 6,000 titles of books. The authors computed four statistics to describe the frequency of occurrence of the words in their corpus: *F* (frequency), *D* (distribution or dispersion), *U* (number of adjusted occurrences per million), and *SFI* (standard frequency index). These same statistics are computed in MANULEX and are described below.

The *Educator's Word Frequency Guide* (Zeno et al., 1995) is based on a corpus of 17 million words—that is, nearly three times the size of Carroll et al.'s (1971) corpus, which is now over 30 years old. It contains 154,941 different word entries. Zeno et al.'s (noncomputerized) norms exceed the earlier studies, not only in the number of words, but also in the number of samples (60,527) and sampled texts, ranging from kindergarten through college. This comprehensiveness and diversity give Zeno et al.'s corpus better coverage of texts in current use across the grades than do any previously published word frequency studies. The guide is divided into four sections. Technical characteristics are described in the first section and then are followed in the second section by an alphabetical list of words with frequencies of 1 or greater. For each word, Carroll et al.'s (1971) statistics (*F*, *D*, *U*, and *SFI*), computed by grade level, are also given. The third section lists words with frequencies less than 1, and the final section presents all the words in the corpus in decreasing order of frequency. Unlike MANULEX, where the grade-level frequency counts are computed by assigning texts to the grade in which they are used, Zeno et al. assigned texts to grade levels on the basis of readability formulas, which determined a range of difficulty values that characterize the materials in each grade (for a discussion of this point in regards to AoA effects, see Zevin & Seidenberg, 2002, note 3).

The CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) is widely used in Europe. For each of three languages (English, Dutch, and German), CELEX provides detailed information on orthography (variations in spelling and hyphenation), phonology (phonetic transcriptions, variations in pronunciation, syllable structure, and primary stress), morphology (derivational and compositional structure and inflectional paradigms), syntax (word class, word-class–specific subcategorizations, and argument structures), and word frequency (summed word and lemma counts based on recent and representative text corpora).

French language. Unlike English, other languages, including French, have a limited number of computerized corpora and lists, or they are still under development. As has been pointed out by Verlinde and Selva (2001), although French lexicographers were among the first to integrate corpus analysis into the dictionary-making process, with the *Trésor de la langue française* project (Imbs, 1971) and its corpus of 170 million words, corpus-based lexicography is not a common practice in contemporary lexicography in France (see above, however, for noncomputerized French lexical databases).

With the FRANTEXT project, French corpus-based lexicography is now in progress. FRANTEXT is an online Web corpus of 3,241 texts, chosen from among 2,330 French literary works and a large group of nonliterary works. The corpus (183 million words) was assembled for the purposes of compiling word occurrences for French dictionary research. FRANTEXT covers a wide variety of aspects of the French language: literary texts (16th–20th centuries), scientific and technical texts (from the 19th and 20th centuries), and regional variations. Texts can be queried by word, sentence, author, title, genre, date, or combinations thereof. Word frequency distribution tables and collocations can be generated for selected words and works.

The BRULEX database (Content, Mousty, & Radeau, 1990) was the first computerized psycholinguistic tool describing the French language. It contains 35,746 entries based on the *Micro Robert* dictionary (Robert, 1986). The frequency counts were taken from the *Trésor de la langue française* for a subcorpus of 23.5 million words in literary texts published between 1919 and 1964.

The LEXIQUE database (New, Pallier, Ferrand, & Matos, 2001) has become the current reference tool in French psycholinguistic research. A subcorpus of texts written since 1950 was extracted from the FRANTEXT corpus (31 million words). The database contains 128,942 word form entries (inflected forms of verbs, nouns, and adjectives) and 54,196 lemma entries. Each entry provides linguistic information, including frequency (per million), gender, number, phonological form, and graphemic and phonemic unicity points. Proper names, symbols, abbreviations, and foreign words were not included. LEXIQUE provides two frequency counts: one based on the 31 million words in the FRANTEXT subcorpus and the other on a Web-based frequency count. For the latter, the words in the FRANTEXT subcorpus were submitted to the search engine FastSearch: The number of pages among 15 million French Web pages where the word was found gives the frequency count. Lemmatization tools were used to obtain the set of lemmas. LEXIQUE 2 (2003) is now available; the number of syllables and the phonetic transcriptions syllabified were corrected by Peereman and Dufour (2003).

Two particular adult databases for psycholinguistic research in French are worth noting. LEXOP (Peereman & Content, 1999) is a computerized lexical database that provides quantitative descriptors of the relationship between the orthography and the phonology of French monosyllabic words. Three main classes of variables are considered: consistency of print-to-sound and sound-to-print associations, frequency of orthography–phonology correspondences, and word neighborhood characteristics. VOCOLEX (Dufour, Peereman, Pallier, & Radeau, 2002) is a lexical database that provides several statistical indexes of phonological similarity among French words (phonological neighbors).

Finally, two recent studies on child language can be mentioned here. Arabia-Guidet, Chevrie-Muller, and Louis (2000) analyzed 118 recent books (100 storybooks, 18 picture books) for preschool children (3–5 years old). Their (noncomputerized) database contains 24,936 words and 8,479 word form entries. No tagging was done to obtain lemmas, and only the most frequent words (254 in storybooks and 101 in picture books) are listed. The frequency count is equal to the number of books in which the word was encountered and, thus, provides an indicator of word use in the books (as in the FastSearch frequency count of LEXIQUE). The NOVLEX database (Lambert & Chesnet, 2001) provides an approximation of the vocabulary of the written materials in use in French elementary schools, but only for third graders. With the help of teachers, the authors selected 38 books (19 third-grade readers and 19 children's storybooks). This corpus gave a total of 417,000 words. The database has 20,600 word form entries and 9,300 lemma entries. For each entry, the frequency of occurrence per 100 million words and the syntactic category are specified.

THE MANULEX DATABASE

The MANULEX database is a grade-based word frequency list extracted from a corpus of first- to fifth-grade readers used in French elementary schools. Four grade levels and, hence, four subcorpora were defined to compute frequencies: first grade (6-year-olds), second grade (7-year-olds), third-to-fifth grades (8- to 11-year-olds), and first-to-fifth grade (i.e., the entire corpus), hereafter called G1, G2, G3–5, and G1–5, respectively. The decision to combine G3, G4, and G5 was based on current research into reading development. Between the first and the fifth grades, children move from emergent literacy to fluent reading by expanding their vocabulary (Adams, 1990). Comprehension processes become more proficient as the child experiences words in specific contexts. The newly acquired word knowledge provides rich semantic associations that can be applied to learning new vocabulary words and so on. In French, more strongly than in English, the most significant changes in reading acquisition occur in the first and the second grades (for longitudinal data, see Sprenger-Charolles, Siegel, Béchennec, & Serniclaes, 2003). As a corollary, whereas G1 and G2 readers contain specific vocabularies, the contents of readers change quantitatively but not qualitatively between G3 and G5, justifying our decision to combine these three reader corpora.

The database contains two lexicons: the word form lexicon and the lemma lexicon, hereafter called the MANULEX word form lexicon (48,886 entries), and the MANULEX lemma lexicon (23,812 entries).

Corpus Sampling

The MANULEX corpus was compiled from reading, spelling, and grammar books by the leading French publishers (see the Appendix for a complete list of the readers and additional information). The readers were selected on the basis of sales for the year 1996. We computed the cumulative sales figures for the set of readers available at each grade and then retained the sample that covered 75% of the sales. So, for each grade, the sample is reasonably representative of printed French materials for schoolchildren 6–11 years of age. This gave us a total of 54 readers: 13 for G1, 13 for G2, and 28 for G3–G5. The readers cover a range of topics, each with a credible amount of data coming from different types of texts (ranging from novels to various kinds of fiction, newspaper reporting, technical writing, poetry, and theater plays) written by different authors from a variety of backgrounds. We did not incorporate other pieces of written materials, such as children's storybooks, because their contents were sufficiently represented in our corpus.

The readers were scanned in their entirety (8,774 pages). Illegible pages were rekeyed. Optical character recognition software was applied to the scanned pages to convert the texts to ASCII format. All areas of the pages were included in the process, except page numbers and some chapter headers.

Tagging and Lemmatization

The term *tagged* (annotated) is used for a corpus that not only contains a sequence of words, but also supplies additional linguistic information associated with particular word forms. The most common linguistic tags are *lemma* (the basic word form) and *grammatical category*. The most reasonable way to tag large corpora is to use computer programs. Cordial Analyseur[3] was chosen here because it performs well under Microsoft Windows. To lemmatize texts, it uses statistical data and explicit rules, along with two types of dictionaries: orthographical dictionaries, which give the lemma of each word (over 117,000 in all), and dictionaries that give grammatical indications (category, gender, and number). The set of syntactic labels used by the analyzer consists of 130 different labels, corresponding to the majority of the morphosyntactic distinctions of French. The statistical and rule-based methods used by Cordial Analyseur perform morphological disambiguation with an acceptable failure rate (1%).

Corpus lemmatization collapses the counts for all inflectional variants of a word into a single lemma count. Other types of inflectional morphology conflated by lemmatization are gender and plural suffixes (e.g., *chat*, *chats*, *chatte*, *chattes* [cat, cats]) and adjective forms (e.g., *corrigé*, *corrigés*, *corrigée*, *corrigées* [corrected]). The rationale for lemmatization was that meaning is usually preserved across the inflectional variants of a lemma, whereas derivational morphological variants are often semantically opaque. Studies on word recognition have demonstrated that lexical processing draws differently upon lemma frequency information (also referred to as stem or summed word form frequency) and word form frequency information. For example, Taft (1979) showed that although *shoe* and *fork* were matched for corpus frequency, *shoe* is recognized faster than *fork* because *shoes* is much more frequent than *forks*. This finding suggests that the basic unit of lexical representation is the lemma, rather than the surface word form. However, Baayen, Dijkstra, and Schreuder (1997) have painted a more complex picture. They found that lexical decision latencies on singular Dutch nouns of differing word form frequency were statistically equivalent when the items had the same lemma frequency. However, this did not hold true for plural nouns, for which word form frequency effects were found. Baayen et al. (1997) proposed that it is more efficient for some morphologically complex words to be stored as wholes, due to orthographic form ambiguity. For instance, in French, some nouns or adjectives (ending in -*ant* or -*ent*) may also correspond to a verb that shares the same stem and, therefore, are ambiguous: *courant* versus *courant* (current vs. running) or *excellent* versus (*ils*) *excellent* (excellent vs. [they] excel).

Frequency Count Computations

The word frequency count is the primary useful output of a corpus (Nation, 2001). As has been pointed out by Nagy and Anderson (1984), the frequency of a word reflects different factors, one of which is conceptual difficulty. In general, a word's frequency can be said to reflect the range of contexts in which the word might appear. However, Francis and Kučera (1982) noted that words are unequally distributed in different types of texts. They pointed out that, unlike high-frequency words, low-frequency words tend to occur in a smaller number of text types—that is, they seem to be context specific. This finding has some important implications here. Indeed, particularly in the first grade, readers differ considerably because editors want to make them attractive and appealing in their design and illustrations. The content is not always selected in the light of teaching aims, and readability seems to be understood differently by the publishers. If a word frequency list should reflect an individual child's exposure to written words, the frequency computed for a word should neither underestimate nor overestimate its occurrences in a corpus of indefinitely large size. For instance, the word *point* (point) was found 276 times in G1, but 242 of these occurrences were in only one reader; the word *papa* (daddy), on the other hand, was found 270 times in G1 and had an even distribution over the set of readers. Clearly, then, a frequency count should take into account the dispersion of occurrences across readers in order to distinguish words recurring in a single context (like *point*) to words recurring in many contexts (like *papa*). Lovelace (1988) emphasized the need for frequency counts based on an index of dispersion. He recommended using Carroll et al.'s (1971) norms instead of Kučera and Francis's (1967) ones, because frequency counts were adjusted in the former to reflect the proportion of contexts in which a word occurred. In MANULEX, the frequency count computations were done in accordance with the methods described by Carroll et al. and, recently, by Zeno et al. (1995). They were computed in the word form and lemma lexicons at all four levels (G1, G2, G3–5, and G1–5). The following description is based on Breland's (1996, p. 97) presentation.

The statistics are as follows.

Frequency. *F* represents the number of times the word type occurs in the corpus.

Dispersion. *D* ranges from .00 to 1.00, on the basis of the dispersion of the frequencies across readers. *D* is equal to .00 when all occurrences of the word are found in a single reader, regardless of the frequency. It is equal to 1.00 if the frequencies are distributed in exactly equal proportions across readers. Values between .00 and 1.00 indicate degrees of dispersion between these extremes. The formula for calculating D is

$$D = \left[\log\left(\sum p_i\right) - \left[\left(\sum p_i \log p_i\right) / \sum p_i\right]\right] / \log(n),$$

where n is the number of readers in the corpus ($n = 13$ in G1, 13 in G2, 28 in G3–5, and 54 in G1–5), i is the reader number ($i = 1, 2, \ldots, n$), and p_i is the frequency of a word in the ith reader, with $p_i \log p_i = 0$ if $p_i = 0$.

Estimated frequency per million words. U is derived from F with an adjustment for D. When $D = 1$, U is computed simply as the frequency per million words. But when $D < 1$, the value of U is adjusted downward. When $D = 0$, U has a minimum value based on the average weighted probability of the word across all the readers. It is believed that U is a better reflection of the true frequency per million that would be found in a corpus of an indefinitely large size, thus permitting direct comparisons with values given by the four subcorpora. The adjustment is done using the following formula:

$$U = \left(1{,}000{,}000 / N\right)\left[FD + \left(1 - D\right) * f_{\min}\right],$$

where N is the total number of words in the corpus (172,348 in G1, 351,024 in G2, 1,386,546 in G3–5, and 1,909,918 in G1–5), F is the frequency of the word in the corpus, D is the index of dispersion, and $f_{\min}$ is $1/N$ times the sum of the products of f_i and s_i, where f_i is the frequency in reader i and s_i is the number of words in that reader.

The standard frequency index. *SFI* is derived directly from U and therefore has some of U's characteristics. The user should find this index to be a simple and convenient way of indicating frequency counts, once it is understood. A word form or a lemma with an *SFI* of 90 is expected to occur once in every 10 words, one with an *SFI* of 80 can be expected to occur once in every 100 words, and so forth. A convenient mental reference point is an *SFI* of 40, the value for a word form or lemma that occurs once in a million words. *SFI* is computed from U by using the formula

$$SFI = 10 * \left[\log_{10}\left(U\right) + 4\right].$$

As an example, we have seen that *point* and *papa* have the same frequency in G1 (276 and 270, respectively). However, they have a different D value (.24 and .79, respectively) and an estimated frequency per million of 507 and 1,270, respectively. Hence, their respective *SFI* values are 67.05 and 71.04.

Description of the Files

The MANULEX database is downloadable at http://www.lpl.univ-aix.fr/lpl/ressources/manulex/ in three formats: ASCII texts (two downloadable lexicon files), Microsoft Excel, and Microsoft Access. When starting to use the database, the user has to choose between the word form lexicon and the lemma lexicon.

The database entries (either word forms or lemmas) vary by syntactic category: noun, proper name, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, determiner, abbreviation, and euphonic string. The database contains four special categories of words that are often excluded from frequency counts: proper names (essentially, first names and countries), compound number words (*dix-huit*, eighteen), abbreviations, and interjections. Unlike some vocabulary researchers, we contend that if a word actually occurs in a corpus, children encounter it in their reading, and it should, therefore, be included in the database (for a similar point of view, see Nagy & Anderson, 1984). The MANULEX word form lexicon yields all possible inflected words reduced to their lemmas in the MANULEX lemma lexicon (the singular for nouns and adjectives, the infinitive for verbs).

For each level (G1, G2, G3–5, and G1–5), after word length and syntactic category (noted NLET and SYNT, respectively), other columns show the frequency of the word in the corpus (F), and Carroll's three computations, D, U, and *SFI* (noted G1 F, G1 D, G1 U, G1 *SFI*, . . . ; G1–5 *SFI*). Empty cells correspond to words not present in a grade level.

Descriptive Statistics

Information about the size of the corpus and the lexicons is presented in Table 1. The corpus contains a total of 8,898,283 characters and a total of 1,925,854 word forms. The database contains only 1,909,918 word forms,

Table 1
Statistics for the MANULEX Corpus and Database

	G1	G2	G3–5	G1–5
Corpus				
Readers (*N*)	13	13	28	54
Characters (including punctuation marks)	765,380	1,605,247	6,527,656	8,898,283
Words (excepting punctuation marks)	174,753	353,841	1,397,260	1,925,854
Database				
Words	172,348	351,024	1,386,546	1,909,918
MANULEX word form entries	11,331	19,009	45,572	48,886
MANULEX lemma entries	6,704	10,400	22,411	23,812
% word forms occurring five or more times	32	31	36	39
% word forms occurring once (hapax)	39	38	33	31
% Lemmas occurring five or more times	43	41	48	50
% Lemmas occurring once (hapax)	29	29	24	23

Table 2
Distribution of Syntactic Categories in the MANULEX Lemma Lexicon (*N* and Percentage)

Syntactic Category	MANULEX Code	Number of Lemma Entries				Percentage			
		G1	G2	G3–5	G1–5	G1	G2	G3–5	G1–5
Noun	NC	3,520	5,149	10,366	10,837	52.5	49.5	46.3	45.5
Proper name	NP	625	1,207	3,780	4,454	9.3	11.6	16.9	18.7
Adjective	ADJ	930	1 689	4 167	4 317	13.9	16.2	18.6	18.1
Verb	VER	1,180	1,751	3,083	3,158	17.6	16.8	13.8	13.3
Adverb	ADV	233	362	713	725	3.5	3.5	3.2	3.0
Interjection	INT	78	89	123	139	1.2	0.9	0.5	0.6
Pronoun	PRO	56	57	61	61	0.8	0.5	0.3	0.3
Preposition	PRE	38	44	52	53	0.6	0.4	0.2	0.2
Abbreviation	ABR	8	11	22	24	0.1	0.1	0.1	0.1
Conjunction	CON	19	21	23	23	0.3	0.2	0.1	0.1
Determiner	DET	14	17	18	18	0.2	0.2	0.1	0.1
Euphonic string	UEUPH	3	3	3	3	0.0	0.0	0.0	0.0
Total		6,704	10,400	22,411	23,812	100	100	100	100

because numerals were removed from the frequency counts. Table 1 also shows that 31% of the word forms and 23% of the lemmas are hapax (one occurrence). Generally, hapax words constitute nearly 50% of the words in a corpus, a ratio which is indicative of a highly varied vocabulary. The present value is in agreement with the need to repeat vocabulary in learning to read.

Table 2 gives the distribution of lemmas by syntactic category at each level (*N* and percentages). Whatever the level, nearly 98% of the lemmas are open-class entries, and half of these are nouns.

Table 3 provides the mean, mode, and percentile values (10, 25, 50, 75, 90) of *SFI* in the MANULEX, NOVLEX, and LEXIQUE databases (lemma lexicons). The statistics are also given for MANULEX with proper names removed from the lexicon, which allows for a more direct comparison with the other databases. The log transformation of *SFI* approximates a symmetric distribution, with the mean close to the median at each level. Consequently, the percentile values may be used in experiments as cutoffs for the selection of high-frequency and low-frequency words (e.g., upper and lower quartiles, respectively). The mean *SFI* reflects the conceptual difficulty of the written word for a schoolchild, with a decrease in the mean and mode indicating an increase in word difficulty. An important drop is observed at level G3–5, where the values approach those of the LEXIQUE database. The significant values (mean, mode, and upper and lower quartiles) become close to those of an adult database when the overall corpus (G1–5) is taken into account. The NOVLEX database (third grade) contains a greater number of frequent words than does the MANULEX G1 lexicon: in G1, the *SFI* mean and mode are 49 and 38, respectively, whereas NOVLEX shows 51 and 44.

Table 4 gives the percentages of nonoverlapping and overlapping lemma entries at each level for the main syntactic categories (open-class items) and for the closed-class items. Among all lemma entries, 51% are nonoverlapping, occurring in the G3–5 subcorpus only (essentially, open-class items). This result shows that it is important to have a lexicon for grades below third grade, because half of the words found in readers for 8-year-olds and above are not present in first- and second-grade readers. The overlapping entries are mainly closed-class items, but 27% of the nouns and 34% of the verbs overlap all three levels. These entries can help in the construction of a new basic vocabulary of the French language (see Lété, 2003).

Table 3
Mean, Mode, and Percentile Values for *SFI* in the MANULEX Lemma, NOVLEX*, and LEXIQUE† Databases

	MANULEX (Proper Names Included)						MANULEX (Proper Names Removed)			
	G1	G2	G3–5	G1–5	NOVLEX	LEXIQUE	G1	G2	G3–5	G1–5
Mean	**48**	**45**	**39**	**37**	**51**	**38**	**49**	**46**	**40**	**39**
Mode	**37**	**36**	**27**	**24**	**44**	**25**	**38**	**36**	**27**	**24**
Minimum	32	29	20	11	44	25	32	29	20	11
Maximum	90	89	89	89	86	88	90	89	89	89
P10	36	33	24	21	44	25	36	33	24	22
P25	**38**	**35**	**27**	**24**	**44**	**30**	**38**	**36**	**28**	**26**
P50	48	44	39	38	49	37	49	45	41	40
P75	**56**	**52**	**48**	**46**	**55**	**45**	**56**	**53**	**49**	**48**
P90	62	59	55	54	60	51	62	59	56	56

Note—Statistics considered significant are shown in boldface. *The lemma lexicon was used. The *SFI* value was computed after calculation of the frequencies per million (field/100). †The FRANTEXT frequencies per million of the lemma lexicon were used (FRANTFREQCUM field); the *SFI* value was computed.

Table 4
Percentage and Mean *SFI* of Nonoverlapping and Overlapping Lemma Entries at Each Level for Open-Class and Closed-Class Items

	Nonoverlapping Entries						Overlapping Entries	
	G1		G2		G3–5		G1–5	
Items	%	*SFI*	%	*SFI*	%	*SFI*	%	*SFI*
Open class								
Noun	1	39	3	36	47	33	27	50
Verb	0	–	2	35	41	33	34	51
Adjective	1	38	2	35	57	33	17	47
Adverb	0	–	1	34	47	33	29	52
Proper name	4	41	10	38	66	30	6	46
Abbreviation	0	–	4	33	46	37	21	48
Interjection	6	35	4	35	25	30	43	49
Closed class								
Conjunction	0	–	0	–	9	48	83	65
Determiner	0	–	0	–	6	25	78	72
Preposition	0	–	2	33	17	40	72	63
Pronoun	0	–	0	–	5	45	90	63
Total	2		4		51		22	

Extensions

In the future, surface word form statistics will be supplied at each level (letter, bigram, trigram, and syllable frequencies). Table 5 provides statistics for the mean number of letters, phonemes, and syllables of open-class entries and of all types of words in the MANULEX word form lexicon. Descriptions of the relationship between orthography and phonology, based on Peereman and Content's (1999) work, are also planned. The computations should provide grapheme–phoneme correspondences (for reading) and phoneme–grapheme correspondences (for spelling). To extend Peereman and Content's adult-based computations on monosyllabic words, the planned study will take into account plurisyllabic words, which cover 90% of the word form entries in MANULEX. Finally, computations of several statistical indexes of phonological similarity between words are planned, as in VOCOLEX (Dufour et al., 2002) for adult French language corpora (for the English language, see De Cara & Goswami, 2002).

Table 5
Mean Number of Letters, Mean Number of Phonemes, and Mean Number of Syllables for Open-Class Entries and All Types of Words in the MANULEX Word Form Lexicon

Syntactic Category		G1	G2	G3–5
Noun	No. of letters	7.0	7.4	8.0
	No. of phonemes	5.0	5.3	5.8
	No. of syllables	2.0	2.2	2.4
Verb	No. of letters	7.5	7.7	8.0
	No. of phonemes	5.8	6.0	6.2
	No. of syllables	2.6	2.7	2.8
Adjective	No. of letters	7.0	7.6	8.3
	No. of phonemes	5.1	5.6	6.2
	No. of syllables	2.2	2.4	2.7
Adverb	No. of letters	7.7	8.9	10.4
	No. of phonemes	5.2	6.2	7.3
	No. of syllables	2.2	2.7	3.2
All types	No. of letters	7.0	7.5	8.0
	No. of phonemes	5.0	5.4	5.8
	No. of syllables	2.1	2.3	2.5

CONCLUSION

With achievement of better experimental control of frequency counts in psycholinguistic, MANULEX could also be used in education for language instruction and vocabulary-grading purposes (see Lété, 2003). MANULEX's grade-level frequency counts should be a useful tool for research on child's vocabulary acquisition and spelling and reading development. By giving French researchers the possibility of manipulating the cumulative frequencies of words, as in Zevin and Seidenberg (2002), the database should contribute to specific debates, such as the roles of AoA and word frequency in visual word recognition. All studies that manipulate or control word frequency in an attempt to gain access to the child mental lexicon should benefit from MANULEX.

REFERENCES

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Arabia-Guidet, C., Chevrie-Muller, C., & Louis, M. (2000). Fréquence d'occurrence des mots dans les livres d'enfants de 3 à 5 ans. *Revue Européenne de Psychologie Appliquée*, **50**, 3-16.

Aristizabal, M. (1938). *Détermination expérimentale du vocabulaire écrit pour servir à l'enregistrement de l'orthographe à l'école primaire*. Louvain: Université de Louvain.

Baayen, R. H., Dijkstra, A. F. J., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory & Language*, **37**, 94-117.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, **7**, 96-99.

Burgess, C., & Livesay, B. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, **30**, 272-277.

Carroll, J. B., Davies, P., & Richman, B. (Eds.) (1971). *The American Heritage word-frequency book*. Boston: Houghton Mifflin.

Catach, N., Jejcic, F., & the HESO group. (1984). *Les listes orthographiques de base du français (LOB): Les mots les plus fréquents et leurs formes fléchies les plus fréquentes*. Paris: Nathan.

Cattell, J. M. (1886). The time taken up by cerebral operations. *Mind*, **11**, 220-242, 377-392, 524-538.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505. [Available: http://www.psych.rl.ac.uk/MRC_Psych_Db.html]

Content, A., Mousty, P., & Radeau. M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'année Psychologique*, **90**, 551-566. [Available: ftp://ftp.ulb.ac.be/pub/packages/psyling/Brulex/]

De Cara, B., & Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments, & Computers*, **34**, 416-423.

Dolby, J. L., Resnikoff, H. L., & MacMurray, F. L. (1963). A tape dictionary for linguistic experiments. In *Proceedings of the American*

Federation of Information Processing Societies: Fall Joint Computer Conference (Vol. 24, pp. 419-423). Baltimore: Spartan Books.

Dottrens, R., & Massarenti, D. (no date). *Vocabulaire fondamental du français*. Neuchâtel: Delachaux & Niestlé.

Dubois, F., & Buyse, R. (1952). Échelle Dubois–Buyse. *Bulletin de la Société Alfred Binet*, No. 405. (Originally published 1940)

Dufour, S., Peereman, R., Pallier, C., & Radeau, M. (2002). VOCOLEX: Une base de données lexicales sur les similarités phonologiques entre les mots français. *L'Année Psychologique*, **102**, 725-746.

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

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavioral Research Methods & Instrumentation*, **12**, 395-427.

Gougenheim, G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental (1º degré)*. Paris: Didier.

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

Imbs, P. (1971). *Dictionnaire des fréquences: Vocabulaire littéraire des XIXe et XXe siècles. I: Table alphabétique. II: Table des fréquences décroissantes*. Nancy: CNRS, Didier.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 1311-1334.

Jones, D. (1963). *Everyman's English pronouncing dictionary*. London: Dent.

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

Käding, J. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: privately published.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associated thesaurus of English and its computer analysis. In A. J. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lambert, E., & Chesnet, D. (2001). NOVLEX: Une base de données lexicales pour les élèves de primaire. *L'Année Psychologique*, **101**, 277-288. [Available: http://www2.mshs.univ-poitiers.fr/novlex/]

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. London: Longman.

Lété, B. (2003). Building the mental lexicon by exposure to print: A corpus-based analysis of French reading books. In P. Bonin (Ed.), *Mental lexicon: Some words to talk about words* (pp. 187-214). Hauppauge, NY: Nova Science.

Lexique 2 (2003). Retrieved from http://www.lexique.org/.

Lovelace, E. A. (1988). On using norms for low-frequency words. *Bulletin of the Psychonomic Society*, **26**, 410-412.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148-197). Hillsdale, NJ: Erlbaum.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, **19**, 304-330.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur Internet: Lexique. *L'Année Psychologique*, **101**, 447-462. [Available: http://www.lexique.org/main/]

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology*, **76**(3, Pt. 2).

Peereman, R., & Content, A. (1999). LEXOP: A lexical database providing orthography–phonology statistics for French monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, **31**, 376-379. [Available: ftp://ftp.ulb.ac.be/pub/packages/psyling/Lexop/]

Peereman, R., & Dufour, S. (2003). Un correctif aux codifications phonétiques de la base de données Lexique. *L'Année Psychologique*, **103**, 103-108. [Available: http://leadserv.u-bourgogne.fr/bases/lexiquecorr/]

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56-115.

Préfontaine, R. R., & Préfontaine, G. C. (1968). *Échelle du vocabulaire oral des enfants de 5 à 8 ans au Canada français*. Montréal: Beauchemin.

Prescott, M. D. A. (1929). Vocabulaire des enfants et des manuels de lecture. *Archives de Psychologie*, **83-84**, 225-274.

Robert, P. (1986). *Dictionnaire du français primordial*. Paris: Dictionnaire le Robert.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523-568.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, **27**, 720-731.

Sprenger-Charolles, L., Siegel, L. S., Béchennec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading and in spelling: A four year longitudinal study. *Journal of Experimental Child Psychology*, **84**, 194-217.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, **7**, 263-272.

Taft, M. (1991). *Reading and the mental lexicon*. Hillsdale, NJ: Erlbaum.

Ters, F., Mayer, G., & Reichenbach, D. (1969). *L'échelle Dubois–Buyse d'orthographe usuelle française*. Neuchâtel: Messeiller.

Thorndike, E. L. (1921). *Teacher's word book*. New York: Columbia Teachers College.

Thorndike, E. L. (1932). *A teacher's word book of 20,000 words*. New York: Columbia Teachers College.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia Teachers College.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.

Vander Beke, G. E. (1935). *French word book*. New York: Macmillan.

Verlinde, S., & Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learner's dictionary. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 594-598). Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language.

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language*, **47**, 1-29.

NOTES

1. From lexique des manuels—that is, the lexicon of readers.

2. The reader will find extensive links on corpora and other computational linguistic resources at http://www-nlp.stanford.edu/links/statnlp.html. The site covers all kinds of linguistic resources available on the World-Wide Web in many languages other than English and French.

3. Copyright Synapse Development, 2001, Version 8.00β.

APPENDIX
List of the Readers in the MANULEX Corpus

Title	Level	US Grade	French Grade	Type	Editor	©	Year Used	No. of Pages	No. of Characters	No. of Words
Au fil des mots	G1	1	CP	REA	Nathan	77	96	126	54,061	12,732
Bien lire à l'école		1	CP/CE1	REA	Nathan	89	96	120	85,959	19,198
Bigoudi et compagnie		1	CP	REA	Nathan	85	95	134	50,133	11,251
C'est à lire		1	CP/CE1	REA	Hachette	93	96	125	70,317	15,673
Daniel et Valérie		1	CP	REA	Nathan	64	96	119	38,531	8,889
Gafi le fantôme		1	CP	REA	Nathan	92	96	178	80,618	19,015
Je lis seul, tu lis seule (autocorrectif)		1	CP	REA	Nathan	89	96	92	20,802	4,598
La ruche aux livres		1	CP/CE1	REA	Hachette	91	97	125	66,137	15,024
Lecture à croquer		1	CP	REA	Magnard	96	96	63	51,179	11,280
Lecture en fête		1	CP	REA	Hachette	93	96	190	80,369	18,063
Lire au CP		1	CP	REA	Nathan	90	96	150	68,966	16,029
Paginaire		1	CP	REA	Hachette	92	95	140	54,547	12,586
Ratus et ses amis		1	CP	REA	Hatier	94	95	125	43,761	10,415
G1 Total		**13**						**1,687**	**765,380**	**174,753**
a.r.t.h.u.r	G2	2	CE1	REA	Nathan	90	96	160	118,246	25,920
C'est à lire		2	CE1	REA	Hachette	91	95	157	123,171	27,355
Eclats de lire		2	CE1	REA	Magnard	90	95	153	109,799	24,140
Gafi le fantôme		2	CE1	REA	Nathan	94	98	157	118,180	26,659
Je lis seul, tu lis seule		2	CE1	REA	Nathan	89	97	92	41,610	9,140
La lecture silencieuse		2	CE1	REA	Nathan	89	96	94	52,264	11,732
La ruche aux livres		2	CE1	REA	Hachette	89	97	157	135,608	30,576
La semaine de français		2	CE1	SPEL	Nathan	88	96	214	203,924	44,813
Langue Française		2	CE1	SPEL	Nathan	95	96	137	136,261	28,902
Le français au CE1		2	CE1	SPEL	Hachette	88	96	245	197,777	42,369
Les 7 clés pour lire et pour écrire		2	CE1	REA	Hatier	92	96	149	114,101	25,243
Paginaire		2	CE1	REA	Hachette	94	95	156	98,863	21,262
Ratus découvre les livres		2	CE1	REA	Hatier	95	96	182	155,443	35,730
G2 Total		**13**						**2,053**	**1,605,247**	**353,841**

APPENDIX (Continued)

Title	Level	US Grade	French Grade	Type	Editor	©	Year Used	No. of Pages	No. of Characters	No. of Words
A la croisée des mots	G3	3	CE2	SPEL	Istra	91	96	220	247,124	52,554
a.r.t.h.u.r		3	CE2	REA	Nathan	89	96	140	142,560	31,097
Bien lire à l'école		3	CE2/CM1	REA	Nathan	87	96	130	167,356	35,336
C'est à lire		3	CE2	REA	Hachette	92	96	189	221,408	48,282
Eclats de lire		3	CE2	REA	Magnard	90	95	183	207,666	45,550
Ixel sait lire		3	CE2	REA	Hachette	94	96	105	109,275	23,138
Je lis seul, tu lis seule		3	CE2	REA	Nathan	90	96	124	90,126	19,311
La lecture silencieuse		3	CE2	REA	Nathan	89	96	194	235,347	52,568
La ruche aux livres		3	CE2	REA	Hachette	90	96	189	200,620	44,290
Langue Française		3	CE2	SPEL	Nathan	95	96	150	209,805	44,441
Les 7 clés pour lire et pour écrire		3	CE2	REA	Hatier	90	95	180	172,012	36,484
G3 Total		**11**						**1,804**	**2,003,299**	**433,051**
a.r.t.h.u.r	G4	4	CM1	REA	Nathan	89	96	125	134,244	28,274
Bien lire à l'école		4	CM1/CM2	REA	Nathan	88	96	130	159,133	33,622
C'est à lire		4	CM1	REA	Hachette	91	94	188	223,893	48,168
Eclats de lire		4	CM1	REA	Magnard	90	95	219	245,949	53,614
La lecture silencieuse (livre 1)		4	CM1	REA	Nathan	88	96	120	153,154	33,636
La ruche aux livres		4	CM1	REA	Hachette	91	96	221	258,157	56,784
La semaine de français		4	CM1	SPEL	Nathan	88	95	280	426,355	88,159
Langue Française		4	CM1	SPEL	Nathan	95	96	200	334,642	69,366
Les 7 clés pour lire et pour écrire		4	CM1	REA	Hatier	89	95	183	199,837	43,324
G4 Total		**9**						**1,666**	**2,135,364**	**454,947**
a.r.t.h.u.r	G5	5	CM2	REA	Nathan	89	96	175	162,442	35,008
C'est à lire		5	CM2	REA	Hachette	92	96	220	316,945	67,795
Eclats de lire		5	CM2	REA	Magnard	90	95	219	264,334	56,708
Je lis seul, tu lis seule (autocorrectif)		5	CM2	REA	Nathan	92	96	80	149,119	32,247
La lecture silencieuse		5	CM2	REA	Nathan	90	96	220	448,315	97,975
La semaine de français		5	CM2	SPEL	Nathan	88	96	270	412,217	87,135
Langue Française		5	CM2	SPEL	Nathan	95	96	200	385,204	78,858
Les 7 clés pour lire et pour écrire		5	CM2	REA	Hatier	88	95	180	250,417	53,536
G5 Total		**8**						**1,564**	**2,388,993**	**509,262**
Total		**54**						**8,774**	**8,898,283**	**1,925,854**

Note—REA, reading books; SPEL, spelling and grammar books.

(Manuscript received April 24, 2003;
accepted for publication July 22, 2003.)