Behavioral Research Methods, Instruments and Computers, 36, 156-166.

MANULEX: A lexical database from French readers.

Bernard LÉTÉ
INRP & Université de Provence

Liliane SPRENGER-CHAROLLES
CNRS & Université de Paris 5

Pascale COLÉ
Université de Savoie

RUNNING HEAD: lexical-database from readers

Correspondance:
Bernard LÉTÉ
Université de Provence
Laboratoire Parole & Langage
29 Avenue Robert Schuman
13621 Aix en Provence Cedex 1
France
email: lete@inrp.fr

Abstract

MANULEX is a Web-accessible database which provides frequency based lists of non-lemmatized and lemmatized words computed from the 1,9 million words of the main French readers. Frequency is provided for four levels: 1st grade (G1), 2nd grade (G2), 3rd to 5th grades (G3-5), and for all grades (G1-5). The frequency computation follows the methods described by Carroll et al. (1971) and Zeno et al. (1995) with 4 indices at each level (F: overall word frequency; D: index of dispersion among the selected readers; U: estimated frequency per 1 million tokens; and, SFI: Standard Frequency Index). The database also provides number of letters and syntactic category information. Other values have been added from LEXIQUE, a database of French adult vocabulary (New & al., 2001): number of phonemes, of syllables, the syllabic units and frequency. MANULEX is intended to provide a useful tool for linguistic analyses and/or to select testing stimuli. It can also be used by researchers in Artificial Intelligence as a source of information on natural language processing to simulate child written language acquisition.

INTRODUCTION

The history of lexicographical studies based on quantitative data is not recent, one of the most often quoted ancestor being Käding (1897) who established a lexical database in order to help those in charge of shorthand writing of political, administrative and commercial speeches in German. It was also in a pragmatic purpose, 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 which was focussed on the establishment of a basic English for language teaching and language diffusion, the core idea being to determine a basic vocabulary, thus necessitating to take into account word frequencies (Thorndike, 1932).

The main goal of these first studies is quite different of that of the recent studies in the same field, that mainly aim to create tools to help linguistic and psycholinguistic researches, the most often quoted tools for American English being the word frequency lists of the Brown Corpus (Kučera & Francis, 1967), the American Heritage Word Frequency Book (Carroll, Davis, & Richman, 1971) and the Thorndike-Lorge Count (Thorndike & Lorge, 1944).

This paper presents MANULEX, the first French linguistic tool providing grade-level frequency based lists (1st to 5th grade) established from the 1,9 million words of the main French readers. It contains 48,886 non-lemmatized entries and 23,812 lemmatized entries. It was compiled with the aim to catch up the works in English language as the latest studies of Zeno, Ivenz, Millard and Duvvuri (1995). It should provide a useful tool for linguistic analyses and/or to select testing stimuli. It should also be used by researchers in artificial intelligence as a source of information on natural language processing to simulate child written French language acquisition. Finally, it should be used in an educational purpose for language instruction, vocabulary grading, syllabus design and materials writing.

Short history of French language lexical databases

Concerning the francophone countries, word frequency tables were established since the beginning of the last century, mainly to help teachers. The first was set-up by Henmon (1924) who wanted to scientifically determine which really were the most usual words and their degree of frequency. This work was mainly based on texts selected in the French literature of the second half of the 19th century. Ten years later, Vander Beke (1935) studied a wider corpus by introducing a proportion of non literary texts, particularly scientific texts and newspaper articles. The main interest of the work were the account of an index of dispersion of words across corpora (a word which appears once in five different corpora being more significant than a word which appears ten times in only one corpus).

The preceding corpora were mainly established from texts for adults. One of the first works including texts written for – and even by – children was presented in the dissertation thesis of Aristizabal (1938) based on 4,100 schoolchildren written productions. The Dubois and Buyse scale (1940) was derived from this work: 3,724 words of the Aristizabal's corpus were dictated to 59,469 primary schoolchildren and classified into 43 steps based on the words correctly spelled. The scale was updated into 40 steps by Ters, Mayer and Reichenbach (1969). In the same line were the study of Dottrens and Massarenti (n.d.) in Switzerland which was based on Prescott's (1929) work, and of Préfontaine and Préfontaine (1968) in Québec who first established a list based on 5 to 8 year-olds' spoken language, list which was after used to select the words for their teaching reading method. The idea of a basic French vocabulary based on spoken corpora was also at the core of the work of Gougenheim, Michéa, Rivenc and Sauvageot (1964) which contains the frequency 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 HESO group (1984) relied on this work, as on two others based on written texts, Imbs (1971) and Juilland, Brodin and Davidovitch (1970), the originality of the latter being to take into account the frequency of lemmatized and non lemmatized words. On these bases, Catach et al. (1984) established a list of the most frequent French words and of their most frequent flexional forms (2,357 entries).

This rapid presentation shows that French researchers in child language development, and French teachers, have poor little tools to do their job. These "databases" are very dated but are still in used because no other alternative exists for child language studies. Researchers essentially rely upon adult language databases (see below). More important, these linguistic materials were extracted from children written productions or adults speech productions. As 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 basically the same way as is assumed for adults, and there is a dramatically greater competence/performance gap for children. As a result, the used of Dubois-Buyse scale or Catach lists to select items for studying, for example, word recognition in French, raises several methodological and theoretical problems. However, these works have opened the way to new French computerized databases which are presented below.

Current computerized language corpora and lexical databases

English language

In English, computerized lexical database were available since the beginning of the 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 (Providence). The corpus consisted of one million words of American English texts printed in 1961 and sampled from 15 different text categories to make it a standard reference. Today, this corpus is considered as small, and slightly dated, but is still in used. 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, designed to represent a wide cross-section of current British English. The BNC is a unique snapshot of the English language, presented so as to render possible almost any kind of computer-based research on language. Leech, Rayson and Wilson (2001) have recently published a word-frequency book derived from the BNC. It includes frequencies for writing and for present-day speech (including everyday conversation).

Some corpora have been compiled in specific lexical databases. The MRC Psycholinguistic Database (Coltheart, 1981) contains 150,837 English words likely to be used in psycholinguistic research and provides information about 26 different linguistic properties. It was established from different sources that took into account most of the factors influencing lexical processing: the Associative Thesaurus (Kiss, Armstrong, Milroy & Piper, 1973), Jones' Pronouncing Dictionary of the English Language (Jones, 1963), Paivio's ratings of the concreteness, imagery and meaningfulness of words (Paivio, Yuille & Madigan, 1968), Gilhooly and Logie's ratings based on age of acquisition, imagery, concreteness, familiarity and ambiguity measures (Gilhooly & Logie, 1980), the Colorado norms which deal with word meaningfulness (Toglia & Battig, 1978), the word frequency counts of Kučera and Francis (1967) and those of Thorndike and Lorge (1944) and the Shorter Oxford English Dictionary database (Dolby, Resnikoff & McMurray, 1963).

The American Heritage Intermediate (AHI: Carroll, Davies & Richman 1971) is based on a survey of US schools. It contains 5,09 million words from publications which were widely read among American schoolchildren aged 7 to 15 years. The set of 86,741 distinct words was created from 500-word samples taken from over 6,000 titles of books. The authors have computed 4 statistics to describe the frequency of occurrence of the words in their corpus. The statistics are F (frequency), D (distribution or dispersion), U (number of adjusted occurrences per million) and SFI (standard frequency index). These statistics are computed in MANULEX and are described below.

The Educator's Word Frequency Guide (EWFG; Zeno & al., 1995) is based on over 17 million tokens and 164,000 types. It is nearly 3 times the size of the corpus in the AHI which is now over 30 years old. The EWFG exceeds the earlier studies not only in number of words, but also in number of samples (60,500) and sampled texts, spanning from kindergarten through college. This comprehensiveness and this diversity give the EWFG corpus better coverage of text in current use across grades than any previously published word frequency study. The guide is divided in four sections. Technical characteristics are described in the first section, followed in section two by an alphabetical list of words with frequencies of 1 or greater. This list includes F, D, U, SFI and frequency by grade-level statistics for each word. Section 3 lists words with frequencies less than 1, and the final section presents the words of the entire corpus in descending order of frequency. In a study on age of acquisition, Zevin and Seidenberg (2002) found that the Zeno et al counts are more closely correlated with latencies than are earlier counts such as those of Kučera and Francis (1967) and CELEX (see below), presumably because of the larger corpus and their inclusion of texts targeted at younger readers.

Another database, which could be used as a foundation for an extension of MANULEX, is the CELEX database (Baayen, Piepenbrock & Gulikers, 1995). For each current language (English, Dutch and German), CELEX provides detailed information on orthography (variations in spelling, hyphenation); phonology (phonetic transcriptions, variations in pronunciation, syllable structure, primary stress); morphology (derivational and compositional structure, inflectional paradigms); syntax (word class, word class-specific subcategorizations, argument structures) and word frequency (summed word and lemma counts, based on recent and representative text corpora). Over the past few years, CELEX data have been successfully used in various types of research and experiments, such as selection of lexical materials for word recognition or word association experiments, study of the mental lexicon through analyses of the distribution of wordlists using several deviation and uniqueness measures, and generation of frequency-based lists of sequences of words, graphemes, phonemes or syllables.

French language

Unlike those for English, computerized corpora and lists for other languages, including French, are limited in number or still under development. As 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 (TLF; Imbs, 1971) and its corpus of 170 million words, corpus-based lexicography is not a common practice in contemporary lexicography in France (see however here above for the non-computerized French lexical databases).

With the FRANTEXT project, French corpus-based lexicography is in progress. FRANTEXT is a web-online corpus of 3,241 texts, chosen among 2,330 works of French literature and a large group of non-literary works. The corpus (183 million words) was assembled for the purpose of compiling word occurrences for French dictionary research. The site was created in 1997 by the INALF (National Institute of the French Language) to present its research programs, particularly its lexicon. FRANTEXT covers all aspects of the French language: literary texts (16^{th}-20^{th} centuries), scientific and technical texts (from the 19^{th}-20^{th} centuries), and regional variations. Texts can be queried by words, sentences, author, title, genre, date or by any combination. Word frequency distribution tables and collocations are generated for selected words and works.

The BRULEX database for French (Content, Mousty & Radeau, 1990) was the first to be machine readable. It contains 35,746 entries based on the Micro Robert dictionary (Robert, 1986). The token frequencies are those of the TLF for a corpus of 23,5 million words of literary texts published between 1919 and 1964.

The LEXIQUE database (New, Pallier, Ferrand & Matos, 2001) is the current reference tool in French psycholinguistic research. A corpus of texts written since 1950 has been extracted from the FRANTEXT corpus (31 million words). The database contains 128,942 wordform entries (inflected forms of verbs, nouns and adjectives) and 54,196 lemma entries. Each entry provides several linguistic informations including frequency (per million), gender, number, phonological form, graphemic and phonemic unicity points. Proper names, symbols, abbreviations and foreign language words have been excluded. LEXIQUE provides two token frequency computations: one based on the 31 million words of the FRANTEXT sub-corpus; the other on a Web frequency count. The wordforms were submitted to the 15 million French Web pages of FastSearch; the number of pages where the token was found gives the token frequency. For the authors, this count provides an estimation of the word usage. Lemmatization tools were used to obtain the set of lemmas.

Finally, two specific adult databases for psycholinguistic research in French can be mentioned. LEXOP (Peereman & Content, 1999) is a computerized lexical database which provides quantitative descriptors of the relation between orthography and phonology for 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, in press) is a lexical database which provides several statistical indexes of phonological similarity between French words (phonological neighbours).

Two recent works on child language can be mentioned. Arabia-Guidet, Chevrie-Muller and Louis (2000) have analyzed 118 recent books (100 storybooks, 18 picture books) for pre-school children (3-5 years old). The corpus contains 24,936 words (8,479 wordform entries). No tagging was made to obtain lemmas. The most frequent words (254 in storybooks and 101 in picture books) are listed. The count was calculated from the number of books where the word was encountered which provides an indice of the word usage in the sample books (as the FastSearch frequency count of LEXIQUE).

The NOVLEX database (Lambert & Chesnet, 2001) provides an estimation of the vocabulary of written material in use in French primary schools, but only for third graders. With the help of teachers, the authors have selected 38 books (19 reading books of third grade and 19 children’s storybooks). The corpus leads to a total of 417,000 words. The database has 20,600 wordform entries and 9,300 lemma entries. For each entry, are provided the frequency of occurrence per 100 million, and the syntactic category.

THE MANULEX DATABASE

The MANULEX database is a word frequency list based on a corpus of readers used in French primary schools, from 1st to 5th grades. It involves three sub-corpora of 1st, 2nd and 3rd to 5th grade and the overall corpus of books (hereafter called G1, G2, G3-5, and G1-5, respectively). It contains two lexicons: the wordform lexicon and the lemma lexicon.

Sampling and Representativeness

McEnery and Wilson (2001) described a modern corpus as any collection of more than one text with four main characteristics: sampling and representativeness, finite size, machine-readable form, and status as standard reference. Two of these are present a-priori in our dataset: it is of finite size and machine readable. Our main concern will be assessing sampling and representativeness.

Our corpus consists in reading, spelling and books published by the French leading publishers (see Appendix for a complete list and additional information). The book selection was made on the basis of the sales for the year 1996. We have computed the cumulative frequency of the sales for the set of books at each grade and we have retained a sample that covered 75% of the sales. So, for each grade, the sample is reasonably representative of printed school French materials for schoolchildren aged 6 to 11 years. This leads to a total of 54 books: 13 in G1, 13 in G2 and 28 in G3-G5. The books cover a range of topic areas, each with a credible size of data coming from different type of texts (from novels to various kinds of fiction; from newspaper reportage to technical writing; from poetry to theater play) written by different authors coming from a variety of backgrounds. This is the reason why we have not incorporated others pieces of written materials, as children's books, because their contents were sufficiently represented in our corpus.
The books were entirely scanned (8,774 pages). The text of the illegible pages was rekeyed. An optical character recognition software was applied to the pages to extract the texts in an ASCII format. All page areas were included in the process except page numbers and some chapter headlines.

Tagging and lemmatization

The term "tagged" (annotated) corpus is used for a corpus which contains not only a sequence of words but also comprises additional information. Typically, this includes linguistic information which is associated with the particular wordforms in the corpus. The most common linguistic tags are lemma (the basic wordform), and the respective grammatical categories.
The most reasonable way to build large annotated corpora is an automatic tagging of the texts by computer programs. However, as pointed out by Ide and Véronis (1998), natural languages display rather complex structure and therefore it is not surprising that attempts to process them by simple deterministic algorithms do not always yield satisfactory results. The result is that the present tagging programs are not able to give fully reliable results and there are many ambiguities in their output.
We have used a tagger that more and more teams use in France, since it performs well under Microsoft® Windows™, and does not require any training. It is commercially distributed, but very cheap for research. It is called Cordial Analyseur®, and is developed by Synapse Development who also developed the Microsoft® Word 2000™ spelling and grammatical tools. The company is one of the founding members of the Natural Language

Understanding Consortium, an international group of linguistic technology experts in five main European languages (English, French, German, Italian, Portuguese, and Spanish). Cordial Analyseur® uses statistical data and explicit rules with two types of dictionaries: orthographical dictionaries which comprise the lemma of each word (more than 117,000 on the whole) and grammatical indications (category, gender and number). For the verbs, a number indicates the type of conjugation. In addition, the analyzer uses another type of dictionary, known as grammatical, which comprises a whole of variables for each word and their directions if they are polysemous. In addition, Cordial Analyseur® has many options, which make possible to regulate in a fine way the regrouping of the words in phrases, to display the lemmas and to obtain various information beyond the simple morpho-syntactic labeling: grammatical functions (subject, object, attribute, etc.) or semantic labels, although obviously this information has not the reliability of the morpho-syntactic labels as pointed out by Valli and Véronis (1999). The set of labels used by Cordial is rather detailed, since it comprises 130 different labels, corresponding to the majority of the morpho-syntactic distinctions of French.

As a result of lemmatizing the corpus, the counts for all inflectional variants of a word are collapsed together into a single lexeme count. Other types of inflectional morphology conflated by lemmatization are gender and plural suffixes (e.g. chat (cat), chats, chatte, chattes), and adjective forms (e.g. corrigé (corrected), corrigés, corrigée, corrigées). Lemmatization was motivated by the observation that meaning is normally preserved across the inflectional variants of a lexeme, whereas derivational morphological variants are often semantically opaque. There is some evidence that lexical processing draws upon lexeme frequency (also referred to as stem or summed wordform frequency) information, in preference to surface wordform statistics. Studies on word recognition demonstrated lexeme frequency to be a better predictor of processing time than simple surface frequency. For example, although shoe and fork are matched for corpus frequency, shoe is recognized faster than fork because shoes is much more frequent than forks (Taft, 1979). This finding suggests that the basic unit of lexical representation is the lexeme, rather the surface wordform. More recently, Baayen, Dijkstra and Schreuder (1997) showed that lexical decision latencies for singular Dutch nouns of differing surface frequency were statistically equivalent when the items were matched for lexeme frequency. However, this was not the case for plural nouns, for which surface frequency effects were found. Baayen et al. (1997) propose 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 verb sharing the same stem, and therefore are ambiguous (courant-courant, current-running; excellent-(ils) excellent, excellent-(they) excel).

Frequency computations

Corpus frequency is an established, robust predictor of word recognition performance. The word frequency effect is one of the earliest empirical observations in cognitive psychology which was made by Cattell (1886). He demonstrate 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 Cattell's pioneering work, word frequency has been a persisting subject of study for investigators concerned with the recognition of words: high frequency words are recognized more quickly and with greater accuracy than low frequency words, whatever the chronometric measure (fixation duration, naming, lexical decision, etc.; see Monsell, 1991, for a review).

Word-frequency counts are the first useful output of a corpus (Nation, 2001). But, as pointed out by Nagy and Anderson (1984), the frequency of a word reflects different factors, one of them being the conceptual difficulty of the word. In general, it might be said that a word's frequency reflects the range of contexts in which the word might appear. Yet, Francis and Kučera (1982) noted that the distribution of words in different type of texts is not equal. They pointed out that unlike high frequency words, low frequency words tend to occur in a smaller number of type of texts. That is, they seem to be context specific. This notion has some important considerations here.Indeed, particularly in 1st grade, there is a great diversity among books because editors want their books to be attractive and appealing in their design and illustrations. The content is not always selected considering the aim of teaching, and the readability seems to be understood differently by the writers. If a word frequency list should reflect individual child's exposure to written words, the frequency computed for a word should not underestimated its apparition in a corpus of indefinitely large size.

In MANULEX, for a given word, are indicated, first, the total number of occurrences in all books and, secondly, its distribution across the different books. This is important in order to ensure that words are not limited to a specific corpus. For instance, in MANULEX, the word point (point) was found 276 times in G1 but with an occurrence of 242 in only one book; whereas the word papa (daddy) was found 270 times in G1 and had an even distribution over the set of books.

For the index computations, we have followed the methods described in Carroll et al. (1971) and used recently by Zeno et al. (1995) in the EWFG (see also Breland, 1996). The Carroll's (1971) statistics were computed in MANULEX (wordform lexicon and lemma lexicon) for the three sub-corpora (G1, G2, and G3-5) and the overall corpus of the books.

F - Frequency, the number of times the word type occurred in the corpus.

D - Dispersion, which can take values from .0000 to 1.000, based on the dispersion of the frequencies over the books. D takes the value .0000 when all occurrences of the word are found in a single book, regardless of the frequency. It would take the value 1.000 if the frequencies were distributed over the books exactly proportionally to the total numbers of tokens (words) in the component lists. Values between .0000 and 1.0000 indicate degrees of dispersion between these extremes. As an example, in the lemma lexicon, "à" has an equal distribution across the 13 books in G1, and thus has a D value of 0.96. "abattre", as another example, occurred only once in G1 and thus has a D value of 0.000; the same word occurred 93 times in G3-5 and had a D value of 0.90, meaning that it occurred with a mean frequency of 4 in each book.

The formula for calculating D may be given as:

$$D = [\log(\Sigma p_i) - [(\Sigma p_i \log p_i) / \Sigma p_i]] / \log(n)$$

where:

> n: amount of books in the corpus (n = 13 in G1; 13 in G2; 28 in G3-5; and 54 in the overall corpus)
>
> i: book number ($i = 1, 2, ..., n$)
>
> p_i: frequency of a token in the ith book, and $p_i \log p_i = 0$ if $p_i = 0$.)

U - the estimated frequency per 1 million tokens, derived from F with an adjustment for D. When D equals 1, U is computed simply as the frequency per 1 million tokens. But when D is less than 1, the value of U is adjusted downward. When D is 0, U has a minimum value based on the average weighted probability of the word type over the books. It is believed that U better reflects the true frequency-per-million that would be found in a corpus of indefinitely large size, thus permitting possible direct comparison to values given by the four corpora.

The adjustment is made by the following formula:

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

where:

N: total number of tokens in the corpus (172,248 in G1; 351,024 in G2; 1,386,546 in G3-5; and 1,909,918 in the overall corpus)

F: frequency of the word in the corpus

D: index of dispersion

f min: 1/N times the sum of the products of fi and si, where fi is the frequency in the book i and si is the number of tokens in the book.

SFI - Standard Frequency Index is derived directly from U and hence has some of the same characteristics as U. It is believed that the user will find this index to be a simple and convenient way of indicating word frequencies, once it is understood. A word type with SFI = 90 would be expected to occur once in every 10 tokens; one with SFI = 80 would be expected to occur once in every 100 tokens, etc. A convenient mental reference point is provided by SFI = 40, the value for a word that would occur once in a million tokens. Each unit of SFI represents an increase of about 25.9% in probability or frequency. SFI is computed from U with the formula:

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

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), and an estimated frequency per 1 million of 507 and 1270, respectively. Hence, their SFI value is 67.05 and 71.04.

Description of the files

The MANULEX database is downloadable at htpp://www… under three formats: ASCII texts (two lexicon files downloadable), Microsoft®Excel™, and Microsoft®Access™.

When starting to use the database, the user first has to choose between two lexicon types hereafter called the MANULEX-wordforms lexicon (48,886 entries) and the MANULEX-lemmas lexicon (23,812 entries).

The database entries (either wordforms or lemmas) vary according to their syntactic category: noun (NC), proper name (NP), verb (VER), adjective (ADJ), adverb (ADV), pronoun (PRO), preposition (PRE), conjunction (CON), interjection (INT), determiner (DET), abbreviation (ABR) and euphonic string (UEUPH). The database contains 4 special categories of words that are often excluded from frequency counts: proper names (essentially surnames and countries), compounds containing numbers (dix-huit), abbreviations and interjections. Unlike some vocabulary researchers, we consider that if a word actually occurs in the corpus, children encounter it in their reading, and we consider this a justifiable operational criterion for including these words in the database (see also Nagy & Anderson, 1984, for a similar point of view).The MANULEX-wordforms lexicon yields all possible inflected words; so, the lexicon contains words like "livre", "livres", "livrer" and so on. In the MANULEX-lemmas lexicon, all inflected wordforms are converted to their lemmas (for nouns and adjectives, the singular; for verbs, the infinitive).

For each sub-corpora (G1, G2, G3-5) and the overall corpus (G1-5), and after the word length and the syntactic category (noted NLET and SYNT, respectively), other columns show the frequency of the word in the corpus and the three Carroll's 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 corpus.

The frequency values of LEXIQUE have been added to give a comparison point of the MANULEX entries with a corpus based on adult language. We have only retained the FRANTEXT frequencies (given per 1 million). FRANTFREQPARM values (FRANTEXT frequencies per million) were added in MANULEX-wordforms; and FRANTFREQCUM values (FRANTEXT cumulative frequencies per million) were added in MANULEX-lemmas (86% and 76% of values recovered, respectively; missing values essentially concern proper names.)

Finally, for each entry recovered, three other fields of LEXIQUE have been added: the number of phonemes, the number of syllables and the phonetic transcriptions syllabified (values corrected by Peereman & Dufour, in press).

Descriptive statistics

The information about the size of the corpus and the lexicons is displayed in Table 1.

INSERT TABLE 1 ABOUT HERE

The corpus provided a total of 8,898,283 characters and a total of 1,925,854 words. The database contained 1,909,918 tokens (digits were removed from the frequency count process). Table 1 also shows that 31% of wordforms and 24% of lemmas are hapax legomena. Generally hapax constitute nearly 50% of the words in a corpus, ratio which is representative of highly varied vocabulary. The present value is in agreement with the need of repeated vocabulary in learning to read.

Table 2 provides the distribution of lemmas by syntactic categories at each level (N and percentages).

INSERT TABLE 2 ABOUT HERE

Whatever the level, half of the lemma entries are concerned with nouns, and near 98% of them are open-class entries.

Table 3 provides the mean, mode and percentile values (10, 25, 50, 75, 90) for SFI in the MANULEX, NOVLEX, and LEXIQUE databases (lemma lexicons). The statistics are also given for MANULEX when proper names are removed from the lexicon, which gives a more direct comparison with the other databases.

INSERT TABLE 3 ABOUT HERE

The log transformation of SFI approximates a symmetric distribution with the mean close to the median at each level. So, in experiments, the percentile values may be used as cut-offs for the selection of high-frequency and low-frequency words (upper and lower quartile, respectively, for example). The mean SFI reflects the conceptual difficulty of written words addressed to schoolchildren, the decreasing of the means and the modes showing increasing vocabulary difficulties. An important drop is observed at the G3-5 level, the values approaching those of the LEXIQUE database. The significant values (mean, mode, upper and lower quartile) become closed to the adult database when the overall corpus (G1-5 level) is taken. The NOVLEX database (3^{rd} grade) contains much more frequent words than MANULEX G1 lexicon: in G1, mean and mode SFI are 49 and 38, respectively, whereas NOVLEX shows 51 and 44.

Table 4 gives the percentages of non-overlapping and overlapping lemma entries at each level, for the main syntactic categories (open-class items) and for the closed-class items.

INSERT TABLE 4 ABOUT HERE

Non-overlapping lemma entries are in the G3-5 sub-corpus, 51% of them (essentially open-class items) being not present in the two other levels. This result shows that it is important to have a lexicon below 3^{rd} grade because half of the words found in books started at 8 year old are not present in 1^{st} and 2^{nd} grade. Overlapping entries are mainly concern with closed-class items, but 27% of the nouns and 34% of the verbs overlap the 3 levels. These entries can help to construct a new basic vocabulary for French language.

Extensions

Computations of surface wordform statistics are planned at each level (letter, bigram, trigram and syllable frequencies). Table 5 provides statistics about mean number of letters, of phonemes, of syllables for open-class entries and for all types of words in MANULEX-wordforms lexicon.

INSERT TABLE 5 ABOUT HERE

Descriptions of relations between orthography and phonology based on the work of Peereman and Content (1999) are planned. The computation should take into account, on the one hand, grapheme-phoneme correspondences (for reading) and, on the other hand, phoneme-grapheme correspondences (for spelling). The study of Peereman and Content only included monosyllabic items. In French, monosyllabic words are very few as provided by our MANULEX count: monosyllabic words are very few (6.70%) and the mean number of syllables of the written words is two. So the Peereman and Content's work needs more in depth analyses.

References

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., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania. Available: http://www.kun.nl/celex/

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

Breland, H.M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. Psychological Science, 7 (2), 96-99.

Carroll, J. B., Davies, P., & Richman, B. (Eds) (1971). The American Heritage Word-Frequency Book. Boston, MA: Houghton Mifflin.

Catach, N., Jejcic, F., & 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.

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

Dolby, J. L., Resnikoff, H. L., & McMurray, F. L. (1963). A tape dictionary for linguistic experiments. Proceedings of the American Federation of information processing societies: Fall Joint Computer Conference, 24, 419-423. Baltimore, MD: Spartan Books.

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

Dubois, F., & Buyse, R. (1940). Échelle Dubois-Buyse. Bulletin de la Société Alfred Binet, 405, 1952.

Dufour, S., Peereman, R., Pallier, C., & Radeau, M. (in press). VOCOLEX: Une base de données lexicales sur les similarités phonologiques entre les mots français [VOCOLEX: A lexical database on phonological similarity between French words]. L'Année Psychologique. Available: http://leadserv.u-bourgogne.fr/bases/vocolex/

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

Gilhooly, K. L., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. Behavioural Research Methods and Instrumentation, 12, 395-427.

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

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

Ide, N., & Véronis, J. (1998). Word Sense Disambiguation: The State of the Art. Computational Linguistics, 24, 1-40.

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

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). Häufigkeiteswörterbuch 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, Rhode Island: Brown University Press.

Lambert, E., & Chesnet, D. (2001). NOVLEX: Une base de données lexicales pour les élèves de primaire [Novlex: A lexical database for elementary grades pupils]. 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.

McEnery, T., & Wilson, A. (2001). Corpus linguistics (2nd edition). Edinburgh: Edinburgh University Press.

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: Lawrence Erlbaum Associates.

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 [A lexical database of contemporary French: LEXIQUE]. L'Année Psychologique, 101, 447-462. Available: http://www.lexique.org/main/

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

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. (in press). Un correctif aux codifications phonétiques de la base de données LEXIQUE [Corrections of phonetic codifications of LEXIQUE database]. L'Année Psychologique. Available: http://leadserv.u-bourgogne.fr/bases/lexiquecorr/.

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.

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

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

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. New York: LEA.

Valli, A., & Véronis, J. (1999). Étiquetage grammatical de corpus oraux: Problèmes et perspectives. Revue Française de Linguistique Appliquée, 4, 113-133.

Vander Beke G. E. (1935). French word book. New-York: The Macmillan Company.

Verlinde, S., & Selva, T. (2001). Corpus-based versus intuition-based lexicography: Defining a word list for a French learners' dictionary. Proceedings of the Corpus Linguistics Conference (pp. 594-598). Lancaster University (UK).

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 and Language, 47, 1-29.

Table 1: Statistics about the MANULEX corpus and database.

	G1	G2	G3-5	G1-5
CORPUS				
Books (N)	13	13	28	54
Characters (including punctuations marks)	765 380	1 605 247	6 527 656	8 898 283
Words (excepted punctuations marks)	174 753	353 841	1 397 260	1 925 854
DATABASE				
MANULEX tokens (different wordforms)	172 348	351 024	1 386 546	1 909 918
MANULEX-wordforms entries	11 331	19 009	45 572	48 886
MANULEX-lemmas entries	6 704	10 400	22 411	23 812
% Wordforms occurring 5 or more	32%	31%	36%	39%
% Wordforms occurring once	39%	38%	33%	31%
% Lemmas occurring 5 or more	43%	41%	48%	50%
% Lemmas occurring once	29%	29%	24%	23%

Table 2: Distribution of the syntactic categories in MANULEX-lemmas lexicon (N and percentages).

Syntactic Category	**Manulex Code**	**Number of Lemma Entries**				**Percentages**			
		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%**

Table 3: Mean, mode and percentile values for SFI in MANULEX-lemmas, NOVLEX[a] and LEXIQUE[b] databases. Significant data are listed in bold italics.

	MANULEX (proper names included)				**NOVLEX**	**LEXIQUE**	**MANULEX (proper names removed)**			
	G1	**G2**	**G3-5**	**G1-5**			**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**
Min	32	29	20	11	44	25	32	29	20	11
Max	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

a: The lemma lexicon was used. The SFI formula was computed after calculation of the frequencies per million (field/100).

b: We have used the FRANTEXT frequencies per million of the overall entries of the lemma database (FRANTFREQCUM field); the SFI formula was computed.

Table 4: Percentages of non-overlapping and overlapping lemma entries at each level with mean SFI, as a function of open-class and close-class items.

	Non-overlapping entries						Overlapping entries	
	G1		G2		G3-5		G1-5	
	%	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%**	

Table 5: Statistics about mean number of letters, mean number of phonemes and mean number of syllables for open-class entries and all types of words in MANULEX-wordforms 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

Appendix: List of reading books used in the present data collection.

Title	Grade	French Grade	Type	Editor	©	Year	Pages	Car.	Words
Au fil des mots	1	CP	LEC	Nathan	77	96	126	54 061	12 732
Bien lire à l'école	1	CP/CE1	LEC	Nathan	89	96	120	85 959	19 198
Bigoudi et compagnie	1	CP	LEC	Nathan	85	95	134	50 133	11 251
C'est à lire	1	CP/CE1	LEC	Hachette	93	96	125	70 317	15 673
Daniel et Valérie	1	CP	LEC	Nathan	64	96	119	38 531	8 889
Gafi le fantôme	1	CP	LEC	Nathan	92	96	178	80 618	19 015
Je lis seul, tu lis seule (autocorrectif)	1	CP	LEC	Nathan	89	96	92	20 802	4 598
La ruche aux livres	1	CP/CE1	LEC	Hachette	91	97	125	66 137	15 024
Lecture à croquer	1	CP	LEC	Magnard	96	96	63	51 179	11 280
Lecture en fête	1	CP	LEC	Hachette	93	96	190	80 369	18 063
Lire au CP	1	CP	LEC	Nathan	90	96	150	68 966	16 029
Paginaire	1	CP	LEC	Hachette	92	95	140	54 547	12 586
Ratus et ses amis	1	CP	LEC	Hatier	94	95	125	43 761	10 415
G1	**13**						**1 687**	**765 380**	**174 753**
a.r.t.h.u.r	2	CE1	LEC	Nathan	90	96	160	118 246	25 920
C'est à lire	2	CE1	LEC	Hachette	91	95	157	123 171	27 355
Eclats de lire	2	CE1	LEC	Magnard	90	95	153	109 799	24 140
Gafi le fantôme	2	CE1	LEC	Nathan	94	98	157	118 180	26 659
Je lis seul, tu lis seule	2	CE1	LEC	Nathan	89	97	92	41 610	9 140
La lecture silencieuse	2	CE1	LEC	Nathan	89	96	94	52 264	11 732
La ruche aux livres	2	CE1	LEC	Hachette	89	97	157	135 608	30 576
La semaine de français	2	CE1	FRAN	Nathan	88	96	214	203 924	44 813
Langue Française	2	CE1	FRAN	Nathan	95	96	137	136 261	28 902
Le français au CE1	2	CE1	FRAN	Hachette	88	96	245	197 777	42 369
Les 7 clés pour lire et pour écrire	2	CE1	LEC	Hatier	92	96	149	114 101	25 243
Paginaire	2	CE1	LEC	Hachette	94	95	156	98 863	21 262
Ratus découvre les livres	2	CE1	LEC	Hatier	95	96	182	155 443	35 730
G2	**13**						**2 053**	**1 605 247**	**353 841**

Title	Grade	French Grade	Type	Editor	©	Year	Pages	Car.	Words
A la croisée des mots	3	CE2	FRAN	Istra	91	96	220	247 124	52 554
a.r.t.h.u.r	3	CE2	LEC	Nathan	89	96	140	142 560	31 097
Bien lire à l'école	3	CE2/CM1	LEC	Nathan	87	96	130	167 356	35 336
C'est à lire	3	CE2	LEC	Hachette	92	96	189	221 408	48 282
Eclats de lire	3	CE2	LEC	Magnard	90	95	183	207 666	45 550
Ixel sait lire	3	CE2	LEC	Hachette	94	96	105	109 275	23 138
Je lis seul, tu lis seule	3	CE2	LEC	Nathan	90	96	124	90 126	19 311
La lecture silencieuse	3	CE2	LEC	Nathan	89	96	194	235 347	52 568
La ruche aux livres	3	CE2	LEC	Hachette	90	96	189	200 620	44 290
Langue Française	3	CE2	FRAN	Nathan	95	96	150	209 805	44 441
Les 7 clés pour lire et pour écrire	3	CE2	LEC	Hatier	90	95	180	172 012	36 484
G3	**11**						**1 804**	**2 003 299**	**433 051**
a.r.t.h.u.r	4	CM1	LEC	Nathan	89	96	125	134 244	28 274
Bien lire à l'école	4	CM1/CM2	LEC	Nathan	88	96	130	159 133	33 622
C'est à lire	4	CM1	LEC	Hachette	91	94	188	223 893	48 168
Eclats de lire	4	CM1	LEC	Magnard	90	95	219	245 949	53 614
La lecture silencieuse (livre 1)	4	CM1	LEC	Nathan	88	96	120	153 154	33 636
La ruche aux livres	4	CM1	LEC	Hachette	91	96	221	258 157	56 784
La semaine de français	4	CM1	FRAN	Nathan	88	95	280	426 355	88 159
Langue Française	4	CM1	FRAN	Nathan	95	96	200	334 642	69 366
Les 7 clés pour lire et pour écrire	4	CM1	LEC	Hatier	89	95	183	199 837	43 324
G4	**9**						**1 666**	**2 135 364**	**454 947**
a.r.t.h.u.r	5	CM2	LEC	Nathan	89	96	175	162 442	35 008
C'est à lire	5	CM2	LEC	Hachette	92	96	220	316 945	67 795
Eclats de lire	5	CM2	LEC	Magnard	90	95	219	264 334	56 708
Je lis seul, tu lis seule (autocorrectif)	5	CM2	LEC	Nathan	92	96	80	149 119	32 247
La lecture silencieuse	5	CM2	LEC	Nathan	90	96	220	448 315	97 975
La semaine de français	5	CM2	FRAN	Nathan	88	96	270	412 217	87 135
Langue Française	5	CM2	FRAN	Nathan	95	96	200	385 204	78 858
Les 7 clés pour lire et pour écrire	5	CM2	LEC	Hatier	88	95	180	250 417	53 536
G5	**8**						**1 564**	**2 388 993**	**509 262**
TOTAL	**54**						**8 774**	**8 898 283**	**1 925 854**