Behavior Research Methods
2007, 39 (3), 593-603

Manulex-infra: Distributional characteristics of grapheme–phoneme mappings, and infralexical and lexical units in child-directed written material

Ronald Peereman
CNRS and Université de Bourgogne, Dijon, France

Bernard Lété
INRP and Université de Lyon, Lyon, France

and

Liliane Sprenger-Charolles
CNRS and Université de Paris V, Paris, France

It is well known that the statistical characteristics of a language, such as word frequency or the consistency of the relationships between orthography and phonology, influence literacy acquisition. Accordingly, linguistic databases play a central role by compiling quantitative and objective estimates about the principal variables that affect reading and writing acquisition. We describe a new set of Web-accessible databases of French orthography whose main characteristic is that they are based on frequency analyses of words occurring in reading books used in the elementary school grades. Quantitative estimates were made for several infralexical variables (syllable, grapheme-to-phoneme mappings, bigrams) and lexical variables (lexical neighborhood, homophony and homography). These analyses should permit quantitative descriptions of the written language in beginning readers, the manipulation and control of variables based on objective data in empirical studies, and the development of instructional methods in keeping with the distributional characteristics of the orthography.

Psycholinguistic studies on reading and writing processes have shown that several characteristics of words, such as their frequency of occurrence, their similarity to other words, and the regularity of the mapping between orthographic and phonological word units, affect performance and acquisition. This often makes it difficult to set up empirical studies due to the fact that each of these characteristics adds additional constraints to stimulus selection. To complicate things further, intercorrelations between variables are frequent, and advances in cognitive modeling and computer simulations (e.g., Ans, Carbonnel, & Valdois, 1998; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Plaut, McClelland, Seidenberg, & Patterson, 1996) suggest that the effects of the variables identified interact. Fortunately, to help in manipulating and controlling variables, detailed linguistic databases are gradually being developed to allow researchers to select adequate stimuli on the basis of objective statistical characteristics. The fact that such methodological tools are now available for different languages is particularly interesting, given that most studies on written-word recognition have been based on English-speaking subjects. Although the findings were initially assumed to be applicable to other alphabetical writing systems, recent observations suggest that the specific characteristics of each writing system modulate both the way words are processed and how literacy is achieved.

The aim of the present paper is to describe a new Web-accessible database, Manulex-infra, specifically designed for studying reading and writing acquisition in French children. The development of Manulex-infra was motivated by the desire to overcome two important limitations of the databases currently available for the study of literacy acquisition. First, as pointed out in the first section of this paper, most current linguistic databases for the French language are either based on adult-directed written material (e.g., Content, Mousty, & Radeau, 1990; New, Pallier, Brysbaert, & Ferrand, 2004; Peereman & Content, 1999) that makes them unsuitable for studying literacy acquisition, or they provide child word-frequency norms that do not reflect current usage. Second, the databases which consider the present-day child's written vocabulary (Lambert & Chesnet, 2001; Lété, Sprenger-Charolles, & Colé, 2004) provide lexical statistics only. However, as developed below (Section 2), reading and writing acquisition are also–and are even mainly so at the beginning–influenced by infralexical variables, particularly grapheme–phoneme consistency. The primary aim of

R. Peereman, ronald.peereman@u-bourgogne.fr

Manulex-infra was thus to develop infralexical statistics, and also to index grapheme–phoneme consistency (GP-consistency index) and phoneme–grapheme consistency (PG-consistency index) using corpora of child-directed French material. The Manulex-infra computations will be described in Section 3, together with some short statistical descriptions of the databases.

1. Lexical Databases for the Study of Literacy Acquisition

Most of the lexical databases developed since the early nineteenth century primarily contain word-frequency counts for written adult-directed or child-directed materials (for a synthesis, see Lété et al., 2004). For example, *The American Heritage Word-Frequency Book* (Carroll, Davies, & Richman, 1971) is based on a corpus of 5.09 million words extracted from publications read by American schoolchildren ages 7 to 15. Similarly, word-frequency databases such as Kučera and Francis (1967) or *Celex* (Baayen, Piepenbrock, & Gulikers, 1995) for the English language, and Imbs (1971; also Content et al., 1990) and *Lexique 2* (New et al., 2004) for the French language, are based on the analyses of large written corpora.

Although such databases have been widely used for experimental purposes, the child word-frequency norms they contain were established about thirty years ago, based on frequency information thought to no longer reflect present day usage. Moreover, the methodology initially applied in these studies is not necessarily adapted to the researcher's needs. To illustrate, the Dubois and Buyse scale (1940/1952), which has been used in some work on French children, does not correspond to word-frequency estimates in texts, but to children's difficulty producing the correct written word form on elementary school dictations. Also, the Gougenheim, Michéa, Rivenc, and Sauvageot (1964) norms, which were included later in the work by Catach, Jejcic, and the HESO group (1984), are based on spoken corpora, not written material.

Fortunately, recent database developments have provided more valuable word-frequency estimates for researchers interested in literacy acquisition. Zeno, Ivenz, Millard, and Duvvuri's (1995) work is based on a corpus of 17 million English words and contains 154,941 different word entries. It surpasses earlier studies (nearly three times the size of the Carroll et al. (1971) corpus) not only in the number of words, but also in the number of samples (60,527) and sampled texts, ranging from kindergarten through college. Another recent database on child language is Masterson, Stuart, Dixon, Desmond, and Lovejoy's (2003) *Children's Printed Word Database*, which is Web-accessible and provides frequency values for words found in books for 5 to 9 year-old children.[1] For the French language, two similar databases for studying literacy acquisition were developed recently. Whereas Novlex (Lambert & Chesnet, 2001) provides frequency norms for words occurring in textbooks for third-grade children, Manulex (Lété et al., 2004) is grade-level-based and includes frequency values for words encountered in French readers designed for first grade, second grade, and third-to-fifth grade.

In spite of the undeniable relevance of the recent word-frequency norms for studying literacy acquisition, word frequency is obviously not the sole factor of performance, so more specialized databases are needed. It is a well-established fact that the performance of beginning and skilled readers (e.g., Balota, Cortese, Sergent-Marshall, & Spieler, 2004; Brand, Rey, Peereman, & Spieler, in preparation; Chateau & Jared, 2003; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995) also depends on other idiosyncratic properties of words, particularly those related to the mapping between orthography and phonology. These observations are consistent with the view that reading processes involve multiple sources of interacting information, both at the lexical and infralexical levels of analysis, and particularly between the orthographic and phonological codes (e.g., Coltheart et al., 2001; McClelland & Rumelhart, 1981; Plaut et al., 1996).

The impact of a writing system's grapho-phonological characteristics on reading and writing performance led several investigators to make use of published word-frequency counts to derive new sets of computations geared to specialized research (e.g., Berndt, D'Autrechy, & Reggia, 1994; De Cara & Goswami, 2002; Gahl, Jurafsky, & Roland, 2004; Jones & Mewhort, 2004; Kessler & Treiman, 1997; Novick & Sherman, 2004; Peereman & Content, 1999; Stanback, 1992; Tamaoka & Makioka, 2004). Unfortunately, there are very few quantitative descriptions of the orthographic and phonological properties of French words that are suitable for studying literacy acquisition. For example, neither Lexique (New et al., 2004) nor Brulex (Content et al., 1990) include orthography-to-phonology statistics for French words. This is also true of the developmental French database Novlex (Lambert & Chesnet, 2001), and the more recent database Manulex on which the present work is based (Lété et al., 2004). When such studies are available, they are not specifically derived from child-directed written material. For example, both Peereman and Content (1999) and Ziegler, Jacobs, and Stone (1996) provided quantitative descriptions of orthography-to-phonology and phonology-to-orthography mappings for words using word-frequency estimates derived from adult-directed written material (Imbs, 1971). An additional limitation is that both of these databases are confined to monosyllabic items, and they index orthography–phonology mappings at the word-rhyme level, not at the grapheme level. While extending the quantitative analyses to polysyllabic words, Lange's (2000) work relied on the same adult frequency norms. Finally, polysyllabic words were also considered in the studies by Véronis (1986) and Catach (1984). However, Véronis's count was based on the Dubois-Buyse scale (1940/1952) which indexes spelling difficulties in young children, and Catach's norms were partially extracted from spoken-language corpora and adult-directed written material. Hence, due to the absence of written word-frequency norms computed from child material, none of the previous studies offer quantitative descriptions of the orthographic and phonological characteristics of words based on today's child-directed written material. The present study was undertaken to fill this gap by providing researchers working

on literacy with objective statistical descriptions based on child-directed corpora. Our aim was thus to develop a companion set of databases to the Manulex frequency norms that will allow experimental studies to manipulate and control several critical infralexical and lexical variables when investigating literacy acquisition in French children. The next section reviews the main variables that have to be considered when studying literacy acquisition, and therefore have to be computed for child-directed written material. Cross-linguistic differences found in empirical studies are emphasized to highlight the critical characteristics of French in literacy acquisition.

2. Main Variables in the Study of Literacy Acquisition: A Short Survey

Psycholinguistic research indicates that word frequency (e.g., Balota & Chumbley, 1984; Frederiksen & Kroll, 1976; Hudson & Bergman, 1985; for a review, see Monsell, 1991) and grapheme–phoneme correspondence (GPC) consistency (e.g., Content, 1991; Content & Peereman, 1992; Jared, 1997; Peereman, 1995; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Ziegler, Perry, & Coltheart, 2003) affect the performance of skilled readers. However, GPC consistency has a greater impact than word frequency at the beginning of reading acquisition, probably because word-frequency effects are item-specific, unlike GPC consistency effects which depend on generalizations based on associations between frequent GPCs. This may be why spelling acquisition has been found to be more difficult than reading acquisition in most language, GPCs (used in reading) being more consistent than phoneme–grapheme correspondences (PGCs, used in spelling), especially in French (Alegria & Mousty, 1996; Leybaert & Content, 1995; Eme & Golder, 2005; Sprenger-Charolles, Siegel, Béchennec, & Serniclaes, 2003; Sprenger-Charolles, Siegel, & Bonnet, 1998; for Spanish: Cuetos, 1989; for German: Wimmer & Hummer, 1990; for English: Bruck & Waters, 1988; Stage & Wagner, 1992).

Strong reliance on GPCs at the onset of literacy also explains why reading acquisition is easier in shallow orthographies than in deep ones (Sprenger-Charolles, 2003; Ziegler & Goswami, 2005). More specifically, cross-linguistic studies indicate that English-speaking children perform reading tasks less well than do children who speak Spanish (e.g., Goswami, Gombert, & Barrera, 1998; Seymour, Aro, & Erskine, 2003), French (e.g., Bruck, Genesee, & Caravolas, 1997; Goswami et al., 1998; Seymour et al., 2003), or German (e.g, Wimmer & Goswami, 1994; Frith, Wimmer, & Landerl, 1998; Seymour et al., 2003). In addition, the gap between English-speaking and non-English-speaking children and adults is larger for pseudowords than for words, that is, when it is not possible to rely on lexical knowledge (for children: Bruck et al., 1997; Frith et al., 1998; Goswami et al., 1998; Seymour et al., 2003; Wimmer & Goswami, 1994; For adults: Paulesu, Démonet, Fazio, McCrory, Chanoine, Brunswick, et al., 2001). These results show that GPC opacity has a strong and long-lasting negative effect on reading that is not only quantitative, but also qualitative.

In addition, the nature of the reading units used by children also hinges on the characteristics of the language. Depending on what language they speak and how well they read, children seem to move rapidly from reliance on surface units (letters) to reliance on more abstract units (graphemes) and larger sublexical units such as onset-rhymes and/or syllables. In English, it is probably because taking word rhymes into account reduces the inconsistencies of GPCs that beginning English readers make greater use of rhyme units (Brown & Deavers, 1999; Goswami et al., 1998; Goswami, Ziegler, Dalton, & Schneider, 2003), unlike beginners in languages with a shallower orthography who mostly rely on GPCs. This is true of Spanish and French (Goswami et al., 1998; Sprenger-Charolles et al., 1998), and even German (Goswami et al., 2003), despite the fact that in German, as in English, there are a greater number of closed than open syllables and thus more opportunity to process rhyme units.

Syllable-based processing also seems to play a more important role in languages where syllable boundaries are clear as they are in French and Spanish for example, unlike English which is often ambisyllabic (Colé, Sprenger-Charolles, Siegel, & Jimenez-Gonzalez, 2004; see also Colé, Magnan, & Grainger, 1999). In addition, reliance on the syllable could depend on the syllabic structure of the language. A finding to this effect was reported by Duncan and Seymour (2003), who noted a significant effect of prosody and syllabic structure in English: the children were more accurate at reading words that had the most frequent English stress pattern (namely, stress on the first syllable; see Cutler & Carter, 1987) and the most frequent English syllabic structure (CVC-CVC). Jimenez and Guzman (2003) reported similar results for bisyllabic words that varied in positional syllabic frequency (PSF), which is the number of times a syllable appears in a particular position in a word (first, second, final). Spanish first and second graders were sensitive to PSF when reading pseudowords aloud, with low-PSF words being read more slowly than high-PFS ones.

Another issue is whether readers process letters or graphemes. This question has been examined especially in French, and in Dutch, where there are many long graphemes, i.e., more than one letter for one phoneme (e.g., in French, Sprenger-Charolles et al., 1998; Sprenger-Charolles, Colé, Béchennec, & Kipffer-Piquard, 2005; in Dutch, Martensen, Maris, & Dijkstra, 2003; in English, Pring, 1981). If the basic unit of the reading process is the grapheme rather than the letter, readers have fewer units to decode and to assemble when the items contain at least one more-than-one-letter grapheme than when they are composed solely of one-letter graphemes. They also have fewer phonemic units to program in order to produce an oral response. Thus, the presence of a multiple-letter grapheme may have a facilitatory effect on reading when the reader relies strongly on GPCs, and when this kind of processing has not yet been automatized as in the youngest children.[2]

In addition to the infralexical characteristics of words, several studies indicate that reading performance is affected by the density of a word's lexical neighborhood, which corresponds to the number of orthographically similar words. Laxon and colleagues (Laxon, Coltheart, & Keating, 1988; Laxon, Masterson, & Moran, 1994) showed that words that are orthographically similar to many other words are read better by children. A similar effect has been observed for French-speaking adults reading words that were GPC-consistent (Dubois-Dunilac, Peereman, & Content, in preparation; Peereman & Content, 1995,1997). The effect was primarily caused by the existence of words that resembled the target word both orthographically and phonologically (phonographic neighborhood density).

It stands out from this short survey that linguistic databases designed for the study of literacy acquisition should provide objective statistics on infralexical variables, especially regarding grapheme-to-phoneme and phoneme-to-grapheme mappings. In addition, the importance of graphemic and syllabic units in word processing also calls for frequency statistics on these units. Finally, it is desirable to supplement these main characteristics with computations on additional variables frequently controlled in empirical studies (such as bigram frequency), using the same word corpora.

Besides infralexical variables, lexical variables have been shown to affect literacy. This is true for the neighborhood characteristics of words, but also for two additional lexical similarity variables, at least in skilled readers. Although their respective roles have not yet been documented in children's reading and writing performance, the existence of heterophonic homographs (different pronunciation but same spelling) and heterographic homophones (same pronunciation but different spelling) has been shown to affect performance in skilled readers (Folk & Morris, 1995; Gottlob, Goldinger, Stone, & Van Orden, 1999; Pexman, Lupker, & Jared, 2001). Finally, findings suggest that word recognition in adults is also partially determined by the location of the uniqueness point (Kwantes & Mewhort, 1999; Lindell, Nicholls, Kwantes, & Castles, 2005; but see Lamberts, 2005). This variable, initially manipulated in auditory word-recognition studies (e.g., Marslen-Wilson, 1984; Radeau & Morais, 1990; Wingfield, Goodglass, & Lindfield, 1997), is defined as the point (phoneme) in the target word where only one lexical candidate remains, given the sequential property of speech. Transposed to reading, the orthographic point of uniqueness corresponds to the serial position of the letter in the word where only one lexical candidate remains, considering the letter strings serially from left to right.

3. Manulex-infra

Our quantitative descriptions of the orthographic and phonological characteristics of French words encountered by children in elementary school were based on the Manulex lexical entries and their corresponding frequency norms (Lété et al., 2004). Our main reason for choosing Manulex instead of Novlex (Lambert & Chesnet, 2001) was the fact that Manulex provides separate lexicon and frequency norms from Grade 1 to Grade 5. This could turn out to be particularly important in computing the orthographic and phonological characteristics of words, because they may depend on the size of the word set encountered by children at the different grade levels.

Manulex is derived from a corpus of 1.9 million words taken from 54 readers used in French primary schools between the first and fifth grades. The readers cover a range of topics, each with an appreciable amount of data coming from different types of texts (from novels to various kinds of fiction, newspaper reporting to technical writing, and poetry to theater plays) written by different authors from a variety of backgrounds. The Manulex corpus also includes grammar books used in the classroom. The database contains two lexicons: the wordform lexicon (48,886 entries), which was used for Manulex-infra, and the lemma lexicon in which words are reduced to their standard forms (23,812 entries). Each lexicon provides a grade-level-based list of words found in first-grade, second-grade, and third-to-fifth grade readers (hereafter called levels G1, G2, G3–5, respectively). A fourth level (G1–5) was generated by combining all readers.

Various frequency computations are available in Lété et al. (2004). These authors used the methods described in Carroll et al. (1971) and Zeno et al. (1995; see also Vander Beke, 1935), with three indexes at each grade level: *F* for overall word frequency, *D* for dispersion across readers, and *U* for estimated frequency per million words, which is derived from *F* with an adjustment for *D*. The grade-based frequency computations were weighted by the index of dispersion across the readers because this allows one to distinguish words recurring in a single reader from words recurring in many readers. This gives a better estimate of the true frequency—i.e., the word usage—that would be found in a corpus of infinite size. For example, the word *point* (point) was found 276 times in G1, 242 of which were in the same reader. The word *papa* (daddy) was found 270 times in G1, with an equal number of occurrences in all G1 readers. Consequently, the two words have *D* values of .24 and .79, respectively, and *U* values of 507 and 1,270, which means that for the same overall frequency *F*, the dispersion index gives an estimated frequency value *U* that is twice as high in one case than the other. In short, the *U* measure combines the number of readers where the word occurs and the word frequency count. It reflects the extent to which words are evenly distributed over multiple readers as opposed to clustered within a few readers. In the present work, Manulex-infra, the token-based computations (frequency weighted) are based on the Manulex *U* index.

All entries in the Manulex-wordform lexicon (Lété et al., 2004) were used for computations except abbreviations, euphonic strings, interjections, and compound entries (entries that contain a space, an apostrophe, or a dash). Because many proper names listed in Manulex have ambiguous or unknown pronunciations, only those with a frequency value of at least .10 in G1–5 levels were considered in the computations. The total number of entries in G1–5 is 45,080. Among these, 10,861 occurred in G1, 18,131 in G2, and 42,422 in G3–5.

Phonological Representations and Phonetic Codes

Most of the planned statistical computations required phonological representations for all orthographic entries occurring on the word list. Because Manulex does not provide such information, phonological codes were added to the database. Some of them were imported from two computerized French lexical databases, *Brulex* (Content et al., 1990), and *Lexique* (New et al., 2004; see also Peereman & Dufour, 2003, for previous phonetic corrections to the *Lexique* database). Some of the phonetic transcriptions were corrected in accordance with the standard French pronunciation listed in the French dictionary *Le Petit Robert* (CD-Rom Version 2.2). Phonetic codes corresponding to words not found in the computerized lexical databases, including those for proper names, were input manually. The phonological representations are based on a system of 16 vowels, 3 glides, and 18 consonants. Note that an additional hatch mark (#) is also used in grapheme–phoneme mappings to indicate a silent grapheme (e.g., the grapheme *t* at the end of a word when not pronounced, as in the word *fort* /fɔR/).

Syllabic and Graphemic Segmentations

Phonological segmentation into syllabic units was required to estimate syllable frequency in each grade. French syllabification is generally unambiguous, since most words include CV (consonant + vowel) syllables. Following the *Maximum Onset Principle* (Clements, 1990), the syllable boundaries in a word such as *paradis* /paRadi/ are located between a vowel and the next consonant (/pa-Radi/). Although CV syllables are frequent in French, many words include intervocalic consonant clusters sur as /bR/ or /st/, as in the words *abris* /abRi/ and *pistil* /pistil/. The presence of such intervocalic consonant clusters causes syllabification to be ambiguous in some cases, and various solutions have been proposed (see Laeufer, 1992). In agreement with previous work on syllabification, the syllabic segmentation principles proposed by Pulgram (1970) were adopted here. Accordingly, syllabic boundaries were located between adjacent consonants except when stops or labio-dental fricatives were followed by liquids (e.g., /bR/, /pl/, /fl/, /fR/, /vl/, /vR/).

Manulex-wordform lexical entries were also segmented into graphemic units in order to compute the frequency and consistency of grapho-phonological mappings. As far as possible, the main principle was to segment the orthographic chains so that each segmented substring corresponds to a single phoneme. The term "grapheme" is thus used here to refer to letter or letter groups that match a phoneme. Note that French includes several multiletter graphemes such as "ou," "an," "un," "in," "eu," "ch," and "gn."

Graphemic segmentation of French words is generally not problematic, although segmentation choices had to be made in some cases. In particular, two broad categories of problems were encountered. On the one hand, the exact limits of the orthographic substring that matched a single phoneme was sometimes ambiguous. For example, The letter "g" is generally pronounced in one of two ways, depending on the vowel that immediately follows it. The pronunciation /g/ occurs in front of the letters "a,""o," and "u" (e.g., *gare*, *golf*, *guide*) whereas the pronunciation /ʒ/ occurs in front of the letters "i" and "e" (*givre*, *gel*). Most of the time, the letter "u" is not pronounced when it appears after the letter "g" but its presence causes the "g" to be pronounced /g/ (*guise* is pronounced /giz/ whereas *gise* is pronounced /ʒiz/). On the other hand, it was sometimes impossible to make each phoneme in a phoneme string correspond to a particular grapheme. Our choices were therefore governed by a second important principle according to which segmentation must highlight inconsistencies in the pronunciation of orthographic strings. The usefulness of allowing graphemic groups to be mapped to more than one phoneme is, for example, particularly obvious for words having *–er* endings. In many words (mainly verbs in the infinitive), "er" corresponds to the phoneme /e/ (*aimer*, *parler*, *viser*), but the pronunciation /ɛR/ is found in some words (*amer*, *fer*, *mer*, *enfer*). Clearly, pronunciation inconsistencies emerge only if the parsing method maintains the graphemic group "er," whatever its pronunciation. In other cases, graphemic groups associated with more than one phoneme are required because no correspondence can be found between letters and individual phonemes. The grapheme "oi" is frequent in French words, and it is generally pronounced /wa/ as in *oiseau* and *noisette*. However, unlike "ui" which that can be broken down into "u" (generally /ɥ/) and "i" (generally /i/), keeping "oi" together is an acceptable solution.

Finally, French is a language that stands apart when it comes to the transcription of word-final morphological marks. A large number of morphological marks that are not pronounced are used in the written language. This is true of derivational marks. For example, the "d" at the end of the French word *lourd* (heavy), from which is derived the word *lourdeur* (heaviness) is silent, whereas the "d" in the English word *kind* (from which *kindness* is derived) is pronounced. In addition, the "s" that signals the plural at the end of a French word (*tables*), is silent, as is the *s* that indicates the second person of verbs (*tu manges* (you eat)), whereas these written letters are pronounced in the English words *tables* or *he/she eats*.[3] The existence of silent letters was taken into account by mapping them to a silent phoneme (represented by a hash mark). A total of 125 graphemic groups was obtained. The interested reader will find more detailed information about the graphemic segmentation principles on the Manulex-infra website.

Computations

The computations fall into two categories: word-length characteristics and grade-level characteristics. The word-length characteristics are the numbers of letters, phonemes, graphemes, and syllables in the word. Contrary to word-length characteristics, grade-level characteristics are function of the word corpus analyzed. They were computed on the four Manulex-wordform lexicons corresponding to the four levels, G1, G2, G3–5, and G1–5, that is, words found in first-grade readers, second-grade readers, third-to-fifth-grade readers, and all readers. There are type-based and token-based computations. Type-based computations are

computations made on each word occurring in a lexicon, whatever its lexical frequency. Thus, a common word like *dans* (in) has the same weight as a word rare like *rang* (rank), despite their large difference in frequency. Token-based computations are computations on each word occurring in a lexicon (word type) weighted by its lexical frequency (taken from the Manulex *U* index). The computations carried out on each of the four lexicons are detailed below.

Association frequencies, GP-consistency, and PG-consistency indexes. The term "association frequencies" refers to the frequency at which a particular grapheme is associated with a particular phoneme, so that grapheme-to-phoneme frequency and phoneme-to-grapheme frequency are identical. The ambiguity of phonological encoding from orthographic input, and the ambiguity of orthographic encoding from phonological input, are generally estimated by consistency indexes. In Manulex-infra, the GP-consistency index is equal to the frequency at which a particular grapheme–phoneme mapping occurs divided by the total frequency of the grapheme, no matter how it is pronounced. For example, the GP-consistency index of the association "ch" → /ʃ/ (as in the word *chat* /ʃa/) is obtained by dividing the frequency of occurrence of the "ch" → /ʃ/ association by the frequency of the grapheme "ch," irrespective of its pronunciation (including /ʃ/, but also /k/ for example, as in *choral* /kɔRal/). The GP-consistency index was then multiplied by 100. Its maximal value (total consistency) is 100. Similarly, the PG-consistency index is equal to the frequency at which a particular phoneme–grapheme mapping occurs, divided by the total frequency of the phoneme multiplied by 100, no matter how the phoneme is spelled.

Consistency can also differ greatly as a function of the serial position of the units in the word. In particular, due to the derivational morphology of French, word endings are often silent, so spelling is less transparent. To better characterize the orthography–phonology mappings of French, frequency and consistency were computed as a function of the relative serial position (initial, middle, final) of the units in the words. Finally, separate tables provide summary statistics on the frequency and consistency values of all associations found in the word corpora. The summary tables should be useful for describing the grapheme–phoneme associations of pseudowords and words not found in the child databases.

Infralexical unit frequencies. Bigrams, biphones, and syllable frequencies were computed for each entry at the four levels as a function of their relative serial position in the word (initial, middle, final). Bigram frequency is the frequency of occurrence of each two-letter sequence in the word list. Transposed to phonology, biphone frequency is the frequency of occurrence of each two-phoneme sequence in the word list. Finally, syllable frequency was computed from the syllabic segmentations of phonological wordforms. Supplementary databases provide summary statistics on bigrams, biphones, and syllable frequencies to allow the user to characterize new stimuli, such as words not occurring in the child databases and pseudowords. Letter, phoneme, and trigram frequency tables are also available.

Orthographic and phonographic neighborhood. Lexical neighborhood density was computed to assess lexical similarities between words. Orthographic neighbors are operationally defined as words that can be generated from the base letter string by a single letter substitution. For example, *race*, *rice*, *rate*, and *rack* are orthographic neighbors of the word *race*. Because orthographic neighborhood density depends on the specific orthographic wordforms known by the children, values at the four levels were computed separately (first grade, second grade, third to fifth grades, all grades).

Phonographic neighborhood density was computed in addition to orthographic neighborhood. Phonographic neighbors are not only orthographically similar, but also phonologically similar to the target word. Phonological similarity between words was estimated by applying the orthographic-neighbor operationalization to phonological forms. Hence, words were considered to be phonologically similar when they differed by a single phoneme. The computation results are incorporated in the main word databases, and the neighbors are listed in separate files along with their frequency.

Homophones and homographs. The number of homophones and the number of homographs for each entry were also computed at the four levels. Again, type-based and token-based computations were performed, the latter by summing homophone or homograph frequencies. While heterophonic homographs are very rare in French, heterographic homophones are numerous, partially due to silent inflectional morphology (e.g., *sans–sang*, *cours–court*, *rat–ras*, where the final consonant is silent). The words entering into the computations are listed in separate files, along with their frequency.

Orthographic uniqueness point. In studies on auditory word recognition, the phonological unicity point is traditionally defined as the serial position of the phoneme (counting from the first phoneme in the word) at which the target word diverges from other lexical candidates. Transposed to orthographic forms, uniqueness point refers to the serial position of the letter (counting from left to right) at which the target word diverges from any other lexical candidates. Orthographic uniqueness point is given for each word in each grade level.

Statistical Descriptions of Manulex-Infra Variables

Our primary aim in providing statistical descriptions of Manulex-infra variables was to allow users to achieve a finer selection of experimental items that takes into account the statistical distributions of the various variables *within the Manulex corpora*. Knowledge of the distribution of variables in a corpus can facilitate selection, so it is useful for the researcher to situate the chosen experimental items relative to the whole corpus for the purposes of estimating their representativeness. An advantage of this approach is that it avoids the use of atypical items that are not representative of the word set in the database. This type of control also appears to be advantageous for cross-linguistic studies, because differences in the methodologies used to elaborate lexical databases (e.g., the

choice of written texts) inevitably lead to variations in the lexical statistics obtained, making cross-linguistic comparisons difficult. Comparisons should be less problematic when they are based on the most representative items of the corpora, which are less likely to be contaminated by initial methodological choices. For this reason, dispersion indexes were calculated for variables considered in Manulex-infra. A full description of the dispersion indexes for each of the variables is available on the Manulex-infra website. A summary is given in Table 1 and briefly outlined below. For the sake of concision, only the by-type mean values are discussed.

In addition to satisfying these methodological considerations, a secondary aim of the statistical descriptions was to make it possible to determine whether infralexical variables, particularly grapheme–phoneme mappings, exhibit distributional disparities across corpora. Obviously, changes in the lexical characteristics of a word set are likely to occur as the size of the set increases. This should be the case for lexical similarity variables such as the number of lexical neighbors or the number of homophonic words. Also, mean word frequency and word length are expected to change as the school grade gets higher, assuming that children are exposed first to short and frequent words. An interesting question, however, is whether the infralexical characteristics of different word sets vary. One possibility is that the mappings between orthography and phonology of words appearing in readers used in the lower grades are less complex. This could be true, for example, if most of the words found in the first-grade readers were purposely selected to minimize grapheme–phoneme inconsistencies. Alternatively, the mean consistency of grapheme–phoneme associations may be similar across grades.

Table 1 shows that mean word frequency decreases with school grade. This shows that the words occurring in the G1 corpus are high-frequency words, and that the words in the corpora corresponding to subsequent grades are less and less frequent. The increase in the size of the word set causes an increase in the number of lexical similarities. Hence, the number of homophonic words rises with the grade level of the corpus. The phonology-to-orthography inconsistency of French is expressed by the high proportion of heterographic homophones (same phonology but different spelling). This proportion evolves only slightly across corpora (82%, 85%, 88% from G1 to G3), suggesting that children learning to read are exposed early on spelling variations of the same sounds. Contrary to heterographic homophones, the proportion of heterophonic homographs (same spelling but different phonology) is very low (less that 0.2% in each of the three grades). As shown in Table 1, the orthographic uniqueness point increases with grade level. This observation is not surprising since word length increases with vocabulary size. More striking is the very small increase in the number of orthographic (and phonographic) neighbors as grade level rises. Indeed, one would have expected neighborhood density to go up with written vocabulary size. The reason

Table 1
Basic Statistics (Type Counts) About the Variables Computed in Manulex-Infra

	G1				G2				G3–5				G1–5			
	Overall	Q25	Q50	Q75	Overall	Q25	Q50	Q75	Overall	Q25	Q50	Q75	Overall	Q25	Q50	Q75
Word frequency (per million)	68.1	0.5	3.8	16.4	42.5	0.2	1.8	8.0	18.8	0.0	0.5	3.0	17.3	0.0	0.3	2.3
No. heterographic homophones	0.9	0.0	1.0	1.0	1.1	0.0	1.0	1.0	1.5	0.0	1.0	2.0	1.6	0.0	1.0	2.0
No. orthographic neighbors	1.1	0.0	0.0	1.0	1.1	0.0	0.0	2.0	1.2	0.0	0.0	2.0	1.3	0.0	1.0	2.0
Orthographic uniqueness point	6.4	5.0	6.0	8.0	6.9	5.0	7.0	8.0	7.6	6.0	8.0	9.0	7.7	6.0	8.0	9.0
Word length (in letters)	7.0	6.0	7.0	8.0	7.4	6.0	7.0	9.0	8.0	6.0	8.0	9.0	8.0	6.0	8.0	9.0
Bigram Frequency																
Initial	136	52	102	195	238	87	170	332	584	218	371	809	619	228	392	875
Middle	354	233	342	460	660	431	641	854	1,660	1,122	1,621	2,124	1,773	1,197	1,735	2,274
Final	376	109	262	463	747	173	446	1,029	1,932	458	1,259	2,232	2,066	487	1,346	2,327
Syllable Frequency																
Initial	106	8	31	113	193	16	54	211	563	42	167	920	596	44	175	958
Middle	28	8	20	40	58	15	40	79	177	45	131	254	187	47	134	269
Final	46	6	19	73	88	10	36	122	253	26	96	395	267	27	103	409
Association Frequency																
Initial	595	336	662	903	996	500	1,110	1,533	2,281	1,037	2,380	3,658	2,431	1,113	2,520	3,904
Middle	1,441	940	1,426	1,861	2,694	1,847	2,695	3,460	7,091	5,227	7,167	8,882	7,584	5,594	7,667	9,507
Final	1,564	480	1,695	2,749	2,741	922	3,232	4,098	6,716	2,165	7,773	11,443	7,181	2,254	8,286	12,236
GPC Consistency																
Initial	96	100	100	100	96	100	100	100	96	100	100	100	96	100	100	100
Middle	80	71	82	93	80	72	81	92	80	72	81	90	80	72	81	90
Final	90	85	98	99	91	85	98	99	92	85	99	99	92	86	99	99
PGC Consistency																
Initial	91	92	99	100	91	90	100	100	91	91	100	100	91	91	100	100
Middle	75	64	76	90	75	65	77	89	76	66	77	88	76	66	77	88
Final	46	33	39	45	45	34	36	45	45	30	40	40	45	30	40	40

why this increase is not observed is most likely due to the fact that the proportion of short words–which have the largest neighborhood density–drops as the school grade increases. For example, the proportion of words of less than 5 letters is 10% in G1, 8% in G2, and 5% in G3–5. A second reason is related to word frequency. Previous corpus analyses on the English language indicate that, on average, frequent words tend to have more orthographic neighbors than less frequent words (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Landauer & Steeter, 1973). The presence of a high proportion of frequent words in G1 may have inflated the average neighborhood density.

Several of the infralexical characteristics presented in Table 1 are also worth considering briefly. As expected, the average bigram frequency rises with the size of the written vocabulary. It also turns out that bigram frequency is higher for word-final bigrams than for initial or middle ones, which means that the words in the corpora are more distinguishable by their first bigram than by their final bigram. A grade-level comparison indicated that the frequency difference between initial and final bigrams increases with vocabulary size. The average frequency also goes up as a function of vocabulary size for syllabic units. However, contrary to what happens for bigrams, syllable frequency is higher in word-initial position than in word-final position. A possible account of this finding is that monosyllabic words were entered into the syllable-frequency computations as having only an initial syllable, thus inflating initial syllable frequency. Further analyses nevertheless indicate that the same result is obtained when monosyllabic words are omitted from the syllable-frequency computations. A more likely reason why initial syllables are more frequent than final ones is that the French language allows for more syllabic structures at the end of a word than at the beginning. Thus, nearly 30% of all multisyllable words have a closed final syllable (ending in a consonant), whereas the initial syllable is closed in only 18% of all words. Regarding grapheme–phoneme associations, frequency naturally increases as a function of the orthographic vocabulary size. Also, as observed for bigrams, the frequency is lower for initial than for final units. The directionality of the orthography–phonology mapping matters, however. As Table 1 indicates, mean consistency is higher for grapheme–phoneme correspondences than for phoneme–grapheme correspondences, an observation that is similar to previous findings based on monosyllabic French words segmented into units larger than the grapheme (onset, nucleus, coda, and rhyme; Peereman & Content, 1998, 1999; Ziegler et al., 1997). On average, the consistency of both grapheme–phoneme and phoneme–grapheme associations does not vary across grades. This finding is similar (although it concerns infralexical properties) to the result mentioned above, that the proportions of heterographic homophones are similar across corpora. It suggests that French children are exposed to highly inconsistent phoneme–grapheme mappings from the very beginning of reading acquisition. Finally, consistency varies with the grapheme position in the word. In particular, phoneme–grapheme consistency is much lower for final graphemes than for initial or middle ones. As discussed above, the high phoneme–grapheme inconsistency is mainly due to the frequent presence of silent derivational marks in word endings.

4. Conclusion

Although lexical databases have a long-standing tradition in educational and psycholinguistic research, the increasing complexity of the various research domains required researchers to control numerous variables in experimental setups and in the interpretation of data. Moreover, two characteristics of the current psycholinguistic approach have prompted the need for new tools. On one hand, testing empirical data against theories is achieved more and more frequently via computational simulations, and it has become necessary to consider structural variables such as regularity or consistency of print-to-sound associations as continuous variables rather than dichotomic ones (e.g., regular words vs. irregular words). In addition, the theories themselves attach more and more importance to the statistical characteristics of the language, as illustrated by studies on word segmentation in young children (e.g., Jusczyk, 1997; Saffran, Newport, & Aslin, 1996).

The elaboration of Manulex-infra meets the need for objective quantitative estimates pertaining to the principal variables thought to influence literacy acquisition and word processing in French-speaking children. The information Manulex-infra provides should be useful for controlling or manipulating variables in experimental studies. The existence of numerous correlations between variables (e.g., between length and neighborhood density) often require controlling several variables likely to vary simultaneously with the manipulated variable. Also, quantitative estimates should be useful for interpreting large performance data through multivariate analyses. Manulex-infra should therefore be a valuable tool for research on reading and writing acquisition. Similarly, by offering the capability of selecting and contrasting word sets that differ along a single dimension, for example grapheme–phoneme consistency, the quantitative information in Manulex-infra can help in designing diagnostic tests to assess reading and writing difficulties in children. Finally, the development of methods for teaching reading and writing can benefit from Manulex-infra's measures of the degree of difficulty of the words that children encounter, as well as the distribution of the grapho-phonological complexities to which they are exposed. Although learning has been reported to be facilitated by a gradual increase in complexity in some cases (e.g., Maxwell, Masters, Kerr & Weedon, 2001), it is still unclear whether a progressive learning approach has real advantages, in general (e.g., Rohde & Plaut, 1999; Elman, 1993) and for the acquisition of literacy in particular. Connectionist models suggest that exposure to the full complexity of orthography–phonology mappings from the start is beneficial to learning and processing new words later (Zevin & Seidenberg, 2002, 2004; Seidenberg & Zevin, in press). Connectionist networks trained on words exhibiting only consistent orthography–phonology associations have trouble in modifying computational weights

for learning new associations. Our distributional analyses of grapheme–phoneme mappings indicate that children are exposed to similar degrees of complexity across school grades, which seams to be an ideal solution according to connectionist simulations. Thus, encountering a variety of grapheme–phoneme associations when starting to read and write may help young children process new words later.

AVAILABILITY

The Manulex-infra databases can be downloaded in various formats (text, Excel, dBase, Access) from leadserv .u-bourgogne.fr/bases/manulex/manulex_infra.

AUTHOR NOTE

This work was supported by grants from the French Ministry of Research (Ecole et Sciences Cognitives) to Arnaud Rey and R.P. and from the Conseil Régional de Bourgogne (Contrat de Plan Etat-Région 2006) to R.P. The authors are grateful to Caroline Calmus for her help with the phonological transcriptions of proper names, and to Vivian Waltz for her assistance with the English. Correspondence concerning this article should be addressed to R. Peereman, Université de Bourgogne, L.E.A.D., Pôle AAFE, Esplanade Erasme, BP 26513, 21065 Dijon Cedex, France (e-mail: ronald.peereman@u-bourgogne.fr).

REFERENCES

Alegria, J., & Mousty, P. (1996). The development of spelling procedures in French-speaking, normal and reading-disabled children: Effects of frequency and lexicality. *Journal of Experimental Child Psychology*, **63**, 312-338.

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review*, **105**, 678-723.

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

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, **133**, 283-316.

Berndt, R. S., D'Autrechy, C. L., & Reggia, J. A. (1994). Functional pronunciation units in English words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 977-991.

Brown, G. D. A., & Deavers, R. P. (1999). Units of analysis in nonword reading: Evidence from children and adults. *Journal of Experimental Child Psychology*, **73**, 208-242.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross linguistic study of early literacy acquisition. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 145-162). Mahwah, NJ: Erlbaum.

Bruck, M., & Waters, G. (1990). An analysis of component spelling and reading skills of good readers–good spellers, good readers–poor spellers and poor readers–poor spellers. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 161-206). San Diego: Academic Press.

Brand, M., Rey, A., Peereman, R., & Spieler, D. (2007). *Syllable frequency effects in disyllabic word reading: A large-scale study*. Manuscript in preparation.

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

Catach, N. (1980). *L'orthographe française: Traité théorique et pratique* [French orthography: Theoretical and practical treatise]. Paris: Nathan.

Catach, N. (1984). *La phonétisation automatique du français: les ambiguïtés de la langue écrite*. Paris: Presse du CNRS.

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.

Chateau, D., & Jared, D. (2003). Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory & Language*, **48**, 255-280.

Clements, G. N. (1990). The role of the sonority cycle in core syllabation. In J. Kingston, & M. E. Beckman (Eds.), *Papers in laboratory phonology: 1. Between the grammar and physics of speech* (pp. 283-333). Cambridge: Cambridge University Press.

Colé, P., Magnan, A., & Grainger, J. (1999). Syllable-sized units in visual word recognition: Evidence from skilled and beginning readers. *Applied Psycholinguistics*, **20**, 507-32.

Colé, P., Sprenger-Charolles, L., Siegel, L., & Jimenez Gonzales, J. E. (2004, June). *Syllables in learning to read in English, French and Spanish*. Paper presented at the SSS–R Congress, Amsterdam.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.

Content, A. (1991). The effect of spelling-to-sound regularity on naming in French. *Psychological Research*, **53**, 3-12.

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.

Content, A., & Peereman, R. (1992). Single and multiple process models of print to sound conversion. In J. Alegria, D. Holender, J. Morais, & M. Radeau (Eds.), *Analytic approaches to human cognition*. Amsterdam: Elsevier.

Cuetos, F. (1989). Lectura y escritura de palabras a traves de la ruta phonologica [Involvement of the phonological reading route in word reading and spelling]. *Infencia y Aprendizaje*, **45**, 71-84.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, **2**, 133-142.

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.

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

Duncan, L. G., & Seymour, P. H. K. (2003). How do children read multisyllabic words? Some preliminary observations. *Journal of Research in Reading*, **26**, 101-120.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, **48**, 71-99.

Eme, E., & Golder, C. (2005). Word-reading and word-spelling styles of French beginners: Do all children learn to read and spell in the same way? *Reading & Writing: An Interdisciplinary Journal*, **18**, 157-188.

Folk, J. R., & Morris, R. K. (1995). Multiple lexical codes in reading: Evidence from eye movements, naming time, and oral reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1412-1429.

Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory & Language*, **32**, 781-804.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 361-379.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Scientific Studies of Reading*, **2**, 31-54.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments, & Computers*, **36**, 432-443.

Goswami, U., Gombert, J. E., & Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French and Spanish. *Applied Psycholinguistics*, **19**, 19-52.

Goswami, U., Ziegler, J. C., Dalton, L., & Schneider, W. (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics*, **24**, 235-247.

Gottlob, L. R., Goldinger, S.D., Stone, G.O., & Van Orden, G.C.

(1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. J*ournal of Experimental Psychology: Human Perceptual Processes*, **25**, 561-574.

GOUGENHEIM, G., MICHÉA, R. RIVENC, P., & SAUVAGEOT, A. (1964). *L'élaboration du français fondamental (1º degré)*. Paris: Didier.

HUDSON, P. T. W., & BERGMAN, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory & Language*, **24**, 46-58.

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.

JARED, D. (1997). Spelling–sound consistency affects the naming of high-frequency words. *Journal of Memory & Language*, **36**, 687-715.

JIMENEZ, J. E., & GUZMAN, R. (2003). The influence of code-oriented versus meaning-oriented approaches to reading instruction on word recognition in the Spanish language. *International Journal of Psychology*, **38**, 65-78.

JONES, M. N., & MEWHORT, D. J. K. (2004). Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior Research Methods, Instruments, & Computers*, **36**, 388-396.

JUSCZYK, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

KESSLER, B. & TREIMAN, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory & Language*, **37**, 295-311.

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

KWANTES, P. J., & MEWHORT, D. J. K. (1999). Evidence for sequential processing in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 376-381.

LAEUFER, C. (1992). Syllabification and resyllabification in French. In C. Laeufer & T. A. Morgan (Eds.), *Theoretical analyses in Romance linguistics* (pp. 18-36). Amsterdam: Benjamins.

LAMBERT, E., & CHESNET, D. (2001). Novlex: Une base de données lexicales pour les élèves de primaire. *L'Année Psychologique*, **101**, 277-288.

LAMBERTS, K. (2005). Interpretation of orthographic uniqueness point effects in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **31**, 14-19.

LANDAUER, T. K., & STREETER, L. A. (1973). Structural differences between common and rare words: Failure or equivalence assumption for theories of word recognition. *Journal of Learning & Verbal Behavior*, **12**, 119-131.

LANGE, M. (2000). *De l'orthographe à la prononciation: Nature des processus de conversion graphème-phonème dans la reconnaissance des mots écrits*. Unpublished doctoral dissertation, Université libre de Bruxelles.

LAXON, V., COLTHEART, V., & KEATING, C. (1988). Children find friendly words friendly too: Words with many orthographic neighbours are easier to read and spell. *British Journal of Educational Psychology*, **58**, 103-119.

LAXON, V., MASTERSON, J., & MORAN, R. (1994). Are children's representations of words distributed? Effects of orthographic neighbourhood size, consistency, and regularity of naming. *Language & Cognitive Processes*, **9**, 1-27.

LE PETIT ROBERT (2001). *Dictionnaires Le Robert*. Electronic version 2.2.

LÉTÉ, B., SPRENGER-CHAROLLES, L., & COLÉ, P. (2004). MANULEX: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers*, **36**, 156-166.

LEYBAERT, J., & CONTENT, A. (1995). Reading and spelling acquisition in two different teaching methods: A test of the independence hypothesis. *Reading & Writing: An Interdisciplinary Journal*, **7**, 65-88.

LINDELL, A. K., NICHOLLS, M. E. R., KWANTES, P. J. K., & CASTLES, A. (2005). Sequential processing in hemispheric word recognition: The impact of initial letter discriminability on the OUP naming effect. *Brain & Language*, **93**, 160-172.

MARSLEN-WILSON, W. D. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 125-150). Hillsdale, NJ: Erlbaum.

MARTENSEN, H., MARIS, E., & DIJKSTRA, T. (2003). Phonological ambiguity and context sensitivity: On sublexical clustering in visual word recognition. *Journal of Memory & Language*, **89**, 375-395.

MASTERSON, J., STUART, M., DIXON, M., DESMOND, L., & LOVEJOY, S. (2003). *The children's printed word data base*. Retrieved July 5, 2006, from www.essex.ac.uk/psychology/cpwd.

MAXWELL, J. P., MASTERS, R. S. W., KERR, E., & WEEDON, E. (2001). The implicit benefit of learning without errors. *Quarterly Journal of Experimental Psychology*, **54A**, 1049-1068.

MCCLELLAND, J. L., & RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception: Part I: An account of basic findings. *Psychological Review*, **88**, 375-407.

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.

NEW, B., PALLIER, C., BRYSBAERT, M., & FERRAND, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**, 516-524.

NOVICK, L. R., & SHERMAN, S. J. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods, Instruments, & Computers*, **36**, 397-401.

PAULESU, E., DÉMONET, J.-F., FAZIO, F., MCCRORY, E., CHANOINE, V., BRUNSWICK, N., ET AL. (2001). Dyslexia, cultural diversity and biological unity. *Science*, **291**, 2165-2167.

PEEREMAN, R. (1995). Naming regular and exception words: Further examination of the effect of phonological dissension among lexical neighbours. *European Journal of Cognitive Psychology*, **7**, 307-330.

PEEREMAN, R., & CONTENT, A. (1995). The neighborhood size effect in naming: Lexical activation or sublexical correspondences? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 409-421.

PEEREMAN, R., & CONTENT, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory & Language*, **37**, 382-421.

PEEREMAN, R., & CONTENT, A. (1998). *Quantitative analyses of orthography to phonology mapping in English and French*. Retrieved July 5, 2006, from homepages.ulb.ac.be/~acontent/OPMapping.html.

PEEREMAN, R., & CONTENT, A. (1999). LexOP: A Lexical database with Orthography–phonology statistics for French monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, **31**, 376-379.

PEEREMAN, R., & DUFOUR, S. (2003). Un correctif aux codifications phonétiques de la base de données LEXIQUE. *L'Année Psychologique*, **103**, 103-108.

PEXMAN, P. M., LUPKER, S. J., & JARED, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 139-156.

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.

POWELL, D., PLAUT, D. C., & FUNNELL, E. (in press). Does the PMSP connectionist model of single word reading learn to read in the same way as a child? *Journal of Research in Reading*.

PRING, L. (1981). Phonological codes and functional spelling units: Reality and implications. *Perception & Psychophysics*, **30**, 573-578.

PULGRAM, E. (1970). *Syllable, word, nexus, cursus*. The Hague: Mouton.

RADEAU, M., & MORAIS J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication*, **9**, 155-164.

REY, A., JACOBS, A. M., SCHIMDT-WEIGAND, F., & ZIEGLER, J. C. (1998). A phoneme effect in visual word recognition. *Cognition*, **68**, 41-50.

ROHDE, D. L. T., & PLAUT, D. C. (1999). Language aquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, **72**, 67-109.

SAFFRAN, J. R., NEWPORT, E. L., & ASLIN, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory & Language*, **35**, 606-621.

SEIDENBERG, M. S., WATERS, G. S., BARNES, M. A., & TANENHAUS, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning & Verbal Behavior*, **23**, 383-404.

SEIDENBERG, M. S., & ZEVIN, J. D. (2006). Connectionist models in de-

velopmental cognitive neuroscience: Critical periods and the paradox of success. In Y. Munakata & M. Johnson (Eds.), *Attention and performance XXI: Processes of change in brain and cognitive development*. Oxford: Oxford University Press.

SEYMOUR, P. H. K., ARO, M., & ERSKINE, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, **94**, 143-174.

SPRENGER-CHAROLLES, L. (2003). Reading acquisition: Cross linguistic data. In T. Nunes & P. Bryant (Eds.), *Handbook of children's literacy* (pp. 43-66). Dordrecht: Kluwer.

SPRENGER-CHAROLLES, L., COLÉ, P., BÉCHENNEC, D., & KIPFFER-PIQUARD, A. (2005). French normative data on reading and related skills: from EVALEC, a new computarized battery of tests. *European Review of Applied Psychology*, **55**, 157-186.

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.

SPRENGER-CHAROLLES, L., SIEGEL, L. S., & BONNET, P. (1998). Phonological mediation and orthographic factors in reading and spelling. *Journal of Experimental Child Psychology*, **68**, 134-155.

STAGE S. A., & WAGNER R. K. (1992). Development of young children's phonological and orthographic knowledge as revealed by their spellings. *Developmental Psychology*, **28**, 287-296.

STANBACK, M. L. (1992). Syllable and rime patterns for teaching reading: Analysis of a frequency-based vocabulary of 17,602 words. *Annals of Dyslexia*, **42**, 196-221.

TAMAOKA, K., & MAKIOKA, S. (2004). Frequency of occurrence for units of phonemes, morae and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers*, **36**, 531-547.

TREIMAN, R., MULLENIX, J., BIJELJAC-BABIC, R., & RICHMOND-WELTY, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, **124**, 107-136.

VANDER BEKE G. E. (1935). *French word book*. New York: Macmillan.

VÉRONIS, J. (1986). Etude quantitative sur le système graphique et phono-graphique du Français. *Cahiers de Psychologie Cognitive*, **6**, 501-531.

WIMMER, H., & GOSWAMI, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, **51**, 91-103.

WIMMER, H., & HUMMER, P. (1990). How German speaking first graders read and spell: Doubts on the importance of the logographic stage. *Applied Psycholinguistics*, **11**, 349-368.

WINGFIELD, A., GOODGLASS, H., LINDFIELD K. C. (1997). Word recognition from acoustic onsets and acoustic offsets: Effects of cohort size and syllabic stress. *Applied Psycholinguistics*, **18**, 85-100.

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.

ZEVIN, J. D., & SEIDENBERG, M. S. (2004). Age of acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, **32**, 31-38.

ZIEGLER, J., & GOSWAMI, U. (2005). Reading acquisition, developmental dyslexia and skilled reading accross languages: A psycholinguistic grain size theory. *Psychological Bulletin*, **13**, 3-29.

ZIEGLER, J. C., JACOBS, A. M., & STONE, G. O. (1996). Statistical analyses of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, & Computers*, **28**, 504-515.

ZIEGLER, J. C., PERRY, C., & COLTHEART, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychonomic Bulletin & Review*, **10**, 947-953.

ZIEGLER, J. C., STONE, G. O., & JACOBS, A. M. (1997). What is the pronunciation for -ough and the spelling for /u/? A data base for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, & Computers*, **29**, 600-618.

NOTES

1. Powell, Plaut, and Funnell (in press) used the database to improve the Plaut et al. (1996) connectionist model of reading. They trained the network on grapheme–phoneme correspondences, and on words found in the database. These modifications caused a sharp improvement in nonword reading, relative to word reading, resulting in a near perfect match with the children's data on this measure.

2. Note that in adults, the presence of multiletter graphemes impairs performance (Rey, Jacobs, Schimdt-Weigand, & Ziegler, 1998; Pring, 1981). However, different results have also been found (e.g., Martensen, Maris, & Dijkstra, 2003).

3. These differences between the spoken and written languages in French are due to its Romance origin. At first, French used inflection marks to indicate, for example, the person of verbs, as in Spanish (cantO, cantAS, . . .), whereas modern French relies on pronouns (je, tu . . .) as in English (I, you, . . .). However, in written modern French there are left-overs from the inflectional system of verbs, like the "s" in tu chantes or the "ent" in ils chantent (which are not pronounced).

(Manuscript received January 20, 2006;
revision accepted for publication June 30, 2006.)