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

Chapter 9

BUILDING THE MENTAL LEXICON BY EXPOSURE TO PRINT: A CORPUS-BASED ANALYSIS OF FRENCH READING BOOKS

Bernard Lété** INRP & CNRS-LPL, France

ABSTRACT

The statistical learning argument states that language input is rich and that it can be learned by representing the probabilistic structure of some of its cues. In line with this conception, an important characteristic of PDP models is that learning is frequency sensitive. The aim of the chapter is to determine whether these views are right in stating that vocabulary acquisition, and hence the lexicon, can be built by the implicit learning of statistical regularities in the linguistic environment. Part 1 explore how the lexicon is built in PDP networks and the ways they account for aspects of vocabulary knowledge. In part 2, through an analysis of lexical probabilities from a large set of French readers (first grade to fifth grade), we try to capture the vocabulary addressed to French school children, and describe how it can develop their reading skills. We conclude that lexical representations can be derived from statistical distribution cues through exposure to print. But there is also no doubt that acquiring lexical knowledge requires a lot of time spent in reading. So, helping children learn about words is certainly the best way to build their lexicon.

Polonius: What do you read my Lord ?
Hamlet: Words, words, words.
-Shakespeare, Hamlet, Act I, Scene 5

Language is very difficult to put into words.
-Voltaire (1694-1778)

Corresponding author: Bernard Lété, Laboratoire Parole et Langage, UMR 6057 CNRS, Université de Provence, 29 Avenue Robert-Schuman, 13621 Aix-en-Provence cedex 1, FRANCE, E-mail: lete@inrp.fr

INTRODUCTION

Words are the building blocks of the language, and it is a commonly accepted fact that the storeroom of word knowledge, the so-called *"mental lexicon", is* the most essential element of language processing. The concept of lexicon was developed by Oldfield (1966) who first proposed the notion of a mental dictionary and raised the question of how information about the meaning of a word is recovered. The adult lexicon is generally viewed as being filled with words that have distinct meanings and independent forms, and which serve as the primitive units of grammatical structure. In this classical view, the lexicon is a passive data structure and the look-up process needed to access the knowledge in it largely constrain its organization in many models (e.g., Forster, 1976; Paap, Newsome, McDonald, & Schvaneveldt, 1982). But, since the pioneering work by McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982), connectionist or parallel distributed processing (henceforth PDP) networks have challenged the classical conception and have contributed to gaining insight into skilled word-recognition processes, including how that skill develops (Seidenberg & McClelland, 1989) and neurological damage that leads to reading disorders (Plaut, 1996; Plaut, McClelland, Seidenberg, & Patterson, 1996). Here, the internal organization of the lexicon is no longer the main issue because the lexicon is contentaddressable, and concepts are an *emergent property* of relations existing in an interconnected lexical network.

This conception, and its implications for human brain processing, contribute to an important debate in cognitive psychology: What is the best way to characterize knowledge representation and processing in the cognitive system in order to account for human language performance? One view (e.g., Pinker, 1991, 1994, 1999) is that language is represented and processed in the form of an explicit set of rules; if an item does not obey the rules (e.g., irregular words in print-to-sound conversions), a separate mechanism is required to handle the exceptions (e.g., the dual route model of word recognition by Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). The alternative view comes from research using PDP networks where computation takes on the form of competitive interactions among neuron-like processing units. Such systems learn by adjusting weights on connections between units in a way that is sensitive to how the statistical structure of the environment influences the behavior of the network. As a result, there is no dichotomy between items that obey the rules and items that do not. As English is a quasi-regular language (Plaut et al., 1996), a network exhibits both rule-like and word-specific behavior without two separate sets of processes being specified. Rather, all items coexist within a single system whose representations and processing reflect the relative degree of consistency in the mappings of different items.

The PDP view is particularly appropriate for symbolizing how knowledge and processes are learned and implemented by the brain. The models make use of a simple but essential idea that all learning is based on a stimulus-response association, and that the mind records the statistical patterns of events in the environment. One method for studying occurrences of such statistical patterns relies on corpus statistics, a dynamic field of research in the past few years. This revival of interest in the study of lexicon acquisition from machine-readable dictionaries is a result of the recent availability of large language corpora and linguistic software which have enabled the computational modeling of a number of psycholinguistic phenomena. Such corpus-based methods in psycholinguistic research have led to new ways of thinking about,

for example, the acquisition of syntactic categories (Redington, Chater, & Finch, 1998), semantic and associative priming (Lund, Burgess, & Atchley, 1995), and vocabulary learning (Landauer & Dumais, 1997). Corpus-based methods are attractive because information about a word's contexts of use can be easily and economically collected for a huge portion of the lexicon, tens of thousands of words --an insurmountable task for conventional methods relying on human intuition.

From the logogen model of word recognition (Morton, 1969) to Plaut et al. (1996) learning in PDP networks and the field of statistical learning, our conceptualization of how word knowledge is represented in our brain has been profoundly changed in the past 30 years. Is there empirical evidence for such frameworks? A crucial variable may be the nature of the written vocabulary that children have experienced. Although no study has presented objective estimates of print exposure (a child's familiarity and experience of written words), it is reasonable to predict that print exposure will be an important determinant of children's word recognition (Cunningham & Stanovich, 1991). In fact, it is often claimed that there are too many words in our written language to be learned through direct teaching. And given that conversations and even television tend to be lexically repetitious (West & Stanovich, 1991), vocabulary expansion must take place mainly through incidental exposure during reading.

A description of the complex process of vocabulary acquisition is beyond the scope of this chapter (see Clark, 1993; Nation, 1990, 2001; Schmitt, 2000; Schmitt & McCarthy, 1997). The aim here is in part to determine whether the PDP view and the statistical learning argument are right in stating that vocabulary acquisition, and hence the lexicon, can be built by the implicit learning of statistical regularities in the linguistic environment. The following main points will be emphasized here. Part 1 will explore how the lexicon is built in PDP networks and the ways this view accounts for aspects of vocabulary knowledge. Then some empirical evidence (essentially supporting the print exposure argument) will be examined to show that children can learn their lexicon inductively. Part 2 will illustrate how a corpusbased approach can help us find out if a child can build his/her lexicon by utilizing statistical regularities in his/her linguistic environment, i.e., through print exposure in a reader.

GENERAL BACKGROUND

Statistical Learning and the Acquisition of the Lexicon

The basic human ability of language understanding --making sense of another person's linguistic input-- does not develop separately from the environment. A growing body of evidence seems to indicate that the encoding of frequency data is an implicit and automatic process, and that infants, children, and adults actually do use frequency data to encode, process, and retrieve linguistic information. For example, several recent studies have stressed that aspects of lexical semantics can be captured by looking at patterns of lexical co-occurrence (Lund & Burgess, 1996; Landauer & Dumais, 1997). In this view, learning the meaning of a word is thought to be at least partially dependent on exposure to the word in its linguistic contexts and there is continuity between how words are acquired and how they are used.

Seidenberg, McDonald and Saffran (2002) depicted statistical learning as giving a powerful framework for approaching two classic learnability problems which strongly limit

the role of language experience. First, it is often posited that languages exhibit properties that must be known innately to the child, because experience provides no evidence of them (this is Chomsky's view, e.g., Chomsky, 1965, 1975). Second, because language input affords unlimited opportunities to make false generalizations, it is often posited that children's minds rapidly converge on the grammar of the language to which they are exposed. As a result, the input is said *"to provide both too little evidence concerning properties of the target language and too much evidence consistent with irrelevant analyses"* (Christiansen, Allen, & Seidenberg, 1998, p. 222).

Contrary to the two above claims, the statistical learning argument states that language input is rich and that it can be learned by representing the probabilistic structure of some of its cues. This argument can be related to the "vocabulary burst" problem. In his criticism of the concept, Bloom (in press) stated that word learning is often said to start slowly but then, at about 16 months of age or when a child has learned about 50 words, things really start to happen. This is called *a word burst, word spurt, vocabulary burst, naming explosion, word explosion,* and so on. From the moment children start uttering their first words around age one, they steadily work on their vocabulary to extend it to about 500 recognizable words when they are two years old. From then on, they will acquire about ten new words a day, working towards an average of 14,000 words in their vocabulary at age six (Carey, 1978) and eventually to the 20,000 to 50,000 words that adult speakers have at their disposal (Aitchison, 1994; Clark, 1993; Nation, 1993). Being faced with the extraordinary task of acquiring all those words in a relatively short period of time, it is only logical that children will apply any means available to them to extend their lexicon, and statistical-cue extraction during extensive reading may be one of the means of doing so.

Christiansen et al. (1998) stressed the PDP view as being the best candidate for representing such statistical cues efficiently because they use more powerful procedures than the behaviorist learning rules, and more interestingly, they are coupled with a theory of knowledge representation that permits the development of abstract, underlying structures. In line with this conception, Seidenberg (1997) stated that PDP networks represent a significant advance over the rule-like position, an important property of PDP algorithms lying in their ability to derive structural regularities from noisy input data, a property that *"is relevant to how the child acquires language under naturalistic conditions" (p. 1600).*

In regards to lexicon acquisition, it will be argued here that, if statistical learning forms the basis of a word's cognitive representation, then the human brain is sensitive to the structure of the environment during language development. As experience with a word accumulates through print exposure, more information about its contexts of use becomes encoded, with a corresponding increase in the ability of the language learner to use the word accurately and appropriately.

PDP Networks and the Lexicon

PDP networks changed the field of cognition in the mid-1980's. In this view, knowledge is represented by shared connections with other knowledge. Network units are called nodes and are processed via statistical properties rather than by the application of rules. They represent a modern rendition of the idea first put forward by Hebb (1949) that complex behaviors may emerge from the operation of aggregations of simple neuronal processing

units. As pointed out by Hulme, Snowling and Quinlan (1991), much of the enthusiasm generated by these models stems from the fact that apparently rule-like behavior may be generated by systems which do not embody explicit representations of these rules.

In line with the statistical learning argument, an important characteristic of PDP models is that learning is frequency-sensitive. These models attempt to capture language regularities for the purpose of improving the performance of the network by estimating the probability distribution of various linguistic units (such as words and sentences). For example, with enough experience, Plaut et al.'s (1996) and Seidenberg and McClelland's (1989) models were able to learn the correct pronunciation of most exception words even without semantic support, thus demonstrating that a network that learns gradually by deriving the statistical structure of orthographic and phonological correspondences was able to read pseudowords as well as regular and irregular words.

As pointed out by Warring (2003), connectionism rests on the assumption that we learn incrementally and through exposure to input by successive trial and error in steps. These steps in the learning process alter associative interconnections by strengthening or weakening of interconnections. The more well-known a piece of word knowledge is, the stronger the interconnection that makes up that part of the word's knowledge. This is consistent with the view that a new word will not be learned completely on the first encounter, but knowledge of that word (such as its pronunciation, spelling, collocates, and so on) will grow incrementally with the number of times the word is read in various contexts.

Where is the lexicon in PDP networks? Rule-like assumptions state that the lexicon is a passive data structure. Words are objects of processing, so their internal representations must be accessed, recognized, and retrieved from permanent storage. The status of words in PDP networks is very different: words are not the objects of processing but input that directly drives the processor. They operate on the network's internal state and move it to another position in the state space. It is more accurate to think of that state as the result of the processing of a word, rather than as a representation of the word itself (Elman, 1995). What the network learns over time is what response it should make to different words, taking context into account. Because words have systematic effects on language behavior, it would not be surprising if all instances of a given word resulted in states which were strongly clustered, or that orthographically, grammatically, or semantically related words produced similar effects on the network (e.g., the neighborhood effect, Grainger, O'Regan, Jacobs, & Segui, 1989, 1992). As in PDP networks at the onset of learning, many lexical entries in the child's lexicon can be considered incomplete. Upon repeated exposure to a particular form in several contexts, new syntactic properties and conceptual representations will be established.

Thus, PDP networks mimic individual experience with print by taking the importance of the current knowledge base into account in acquiring new information. One line of empirical research which provides strong support for this view is Stanovich and colleagues' *print exposure* argument regarding individual differences in reading abilities. They demonstrated that differences in reading volume can account for a considerable portion of the variance in knowledge and vocabulary acquisition among children (Cunningham & Stanovich, 1991) and young adults (Stanovich & Cunningham, 1992, 1993; West & Stanovich, 1991).

Building the Lexicon by Print Exposure

Children who read more are better readers than children who read less, and as a consequence, to build their lexicon, children need to have acquired an adequate skill with print. Measures *of "print exposure" or "reading volume",* terms used to refer to a child's familiarity and experience with written words, account for significant variations in word recognition. This classic evidence fits well with PDP networks, where learning is shaped by environmental experience, i.e. the range of words encountered by the network and the number of times each has been encountered.

There is a line in the book of Matthew that says, *"For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath"* (XXV:29). Loosely translated, what this line means is that *"the rich get richer and the poor get poorer"*. This quote was the inspiration for naming a phenomenon *"The Matthew Effect"* that has been observed many times in reading research. The term was first coined by Walberg and Tsai (1983), but Stanovich (1986) popularized the concept and provided a thorough understanding of it based on years of research (e.g., Cipielewski & Stanovich, 1992; Cunningham & Stanovich, 1991, 1993; Stanovich, 1986, 1993, 2000; Stanovich & Cunningham, 1992, 1993; Stanovich & West, 1989; Stanovich, West, Cunningham, Cipielewski, & Siddiqui, 1996; West & Stanovich, 1991). In line with the *Matthew Effect*, some children begin school with the foundational skills that lead to reading success already under their belt, whereas other children are not so "lucky" and begin school with very little experience with print. But, what Stanovich and colleagues stressed is that in kindergarten and first grade, the gap is still surmountable, and teachers can help all children gain the necessary skills for reading success. Beyond the first grade, however, the gap becomes increasingly large. By the fourth grade, helping children gain these foundational skills is time-consuming and usually very frustrating for the child.

Stanovich (1986) pointed out that the *Matthew Effect* concept stems from the finding that individuals who have advantageous early educational experiences are able to utilize new educational experiences more efficiently. Stemberg (1985, cited in Stanovich, 1986) articulated this point in the context of vocabulary acquisition: *"Vocabulary is not only affected by operations of components, it affects their operations as well. If one grows up in a household that encourages exposure to words, then one's vocabulary may well be greater, which in turn may lead to a superior learning and performance on other kinds of tasks that require vocabulary" (p.* 123). According to Stanovich (1986), this difficulty can be explained by an interaction between vocabulary knowledge and reading ability. Children who are good readers encounter greater amounts of text than do poor readers. Thus, better readers are exposed to more words and are able to access a greater number of meanings from context than their classmates who are experiencing reading difficulties. They learn more words meanings incidentally, making further reading easier. On the other hand, struggling readers experience a negative cycle. They begin with a smaller reading vocabulary, are exposed to less text, and encounter fewer words. In addition, it is likely that they will be less able to make efficient use of context to derive the meanings of new words, thereby minimizing their ability to expand their reading vocabulary incidentally. This results in an ever-widening gap between good and poor readers.

In line with Stanovich's work, most theorists agree that the bulk of vocabulary growth during a child's lifetime occurs indirectly through language exposure rather than through

direct teaching (Nagy & Anderson, 1984; Nagy, Herman, & Anderson, 1985; Sternberg, 1985, 1987). Furthermore, many researchers are convinced that reading volume, rather than speech, is the prime contributor to individual differences in children's vocabularies (Hayes & Ahrens, 1988; Nagy & Anderson, 1984; Nagy & Herman, 1987; Stanovich, 1986). Indeed, one empirical reason for believing that a high reading volume is a particularly effective means way of expanding a child's vocabulary, which is derived from differences in the statistical distributions of words that have been found between print and speech (see below). Studies of implicit vocabulary acquisition have shown that learning through extensive reading is not only possible, but is almost certainly the means by which native speakers acquire the majority of their vocabulary. In order to infer meaning, however, the reader must know approximately 95% of the running words in the text (Laufer, 1989; Nation, 1990; Parry, 1991).

But what does it mean to "know" a word? Nation (1990) proposes the following list of different kinds of knowledge that a person must master in order to know a word: its meaning(s), its written form, its spoken form, its grammatical behavior, its collocations, its associations, and its frequency. These different types of word knowledge are not necessarily learned at the same time, but in a gradual manner, and some may develop later than others and at different rates. From this perspective, vocabulary acquisition is incremental; words are not discovered abruptly, but gradually become more and more likely as all this word knowledge accumulates. Thus, unless the word is completely unknown or fully acquired, the different kinds of word knowledge will exist at various degrees of mastery (Schmitt, 2000). Here, we are plainly in a PDP network view, where the strength of the interconnections reflects the relative knowledge an individual has about an item of vocabulary. Word representations emerge as a natural outcome of the learning process. Each network of knowledge is connected to many other networks, and the richer the network of associations, the greater the chances of comprehension (Waring, 2003).

Implications for Vocabulary Teaching

Reading is a difficult linguistic task, and it requires instructional support. Exactly how much instruction is necessary to get the child started on the path to literacy, we do not know (for an up-to-date synthesis on reading acquisition and reading teaching, see Stanovich, 2000 and Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). Almost every classroom teacher recognizes the need to teach vocabulary, and most teachers do so because vocabulary knowledge is significantly related to reading comprehension, decoding, spelling, and school achievement. The teacher's responsibility lies mainly in setting up exposure to language in a vivid way and encouraging reading of material that children care about. Motivation to read is related to the expansion of reading power because, to become effective readers, children must have both the skill and the will to read. Gambrell (1996) showed that classroom environments that foster reading motivation are characterized by a teacher who is a reading model and by a book-rich reading environment, opportunities for choice, familiarity with books, social interaction about books, and literacy-related incentives that reflect the value of reading.

Useful lexical knowledge that can be transferred to the comprehension of novel text seems to require coming across new words in rich natural contexts (Mezynski, 1983). However, given the haphazard nature of acquisition from context, this type of learning requires time (Nagy et al., 1985). The goal for vocabulary development is to insure that

students are able to apply their knowledge of words to appropriate situations, and to increase and enrich their knowledge through independent encounters with words. Research indicates that the best way to help students reach this goal is to add specific words to their lexicon, enhance skills that promote the independent learning of words, and provide opportunities in which words can be learned (Beck & McKeown, 1991).

A Corpus-Based Approach to Studying Print Exposure in French Primary School Children

Cobb (1997) pointed out that in classical Chomskyan linguistics, empirical evidence played a relatively minor role in describing language. In Chomsky's view, the data to be explained by a linguistic theory are native speakers' intuitions about their language, not the infinite minutiae of the language itself. Language is learnable because the system is in the head, not in the ambient linguistic evidence. However, corpus linguistics has established that the system is not in the head only, and that language learning relies more on linguistic evidence than Chomsky allowed.

The corpus-based approach has been used in speech segmentation studies to resolve the so-called bootstrapping problem of vocabulary acquisition in infants (Batchelder, 1997; Cairns, Shillcock, Chater, & Levy, 1997). Research on vocabulary acquisition includes Nagy and Anderson's (1984) study on the number of distinct words in printed school English. This study was conducted on a 7,260 word sample of the Carroll, Davies and Richman's (1971) *Word Frequency Book,* which is based on a corpus of 5.08 million words of running text from many types of materials in used in schools. Projecting from this sample to the total vocabulary of school English, they estimated that there were about 88,533 distinct word families. For Nagy and Anderson, these word families result in a total volume of nearly 500,000 graphically distinct words, including proper names; and roughly half of these 500,000 words occur one or fewer times in a billion words of text. A child's vocabulary size approximately doubles between grades 3 and 7. Finally, Nagy and Anderson estimated that an average student in grades 3 through 12 is likely to learn approximately 3,000 new vocabulary words each year, assuming he or she reads between 500,000 and a million running words of text a school year. Between grades 1 and 3, it is estimated that economically disadvantaged students' vocabularies increase by about 3,000 words per year and middle-class students' vocabularies, by about 5,000 words per year.

Given Nagy and Anderson's (1984) data, are there too many words in printed school materials to be learned through direct teaching, and must vocabulary expansion take place mainly through incidental exposure during reading? Through an analysis of lexical probabilities and some corpus statistics from a large set of French readers (first grade to fifth grade), this chapter will next try to capture and describe the vocabulary addressed to French school children, and describe how it can develop their reading skills. This information should help us answer the above question.

THE LEXICON ADDRESSED TO A CHILD IN FRENCH READERS

Number of Words in French Readers

To do their job, French researchers interested in language acquisition essentially rely upon adult language databases *(Brulex:* Content, Mousty, & Radeau, 1990; *Lexique:* New, Pallier, Ferrand, & Matos, 2001) or upon linguistic materials extracted from children's written productions *(Dubois & Buyse's scale,* 1940) or adults' speech productions (Gougenheim, Michéa, Rivenc, & Sauvageot, 1964).

Manulex[1] (Lété, Sprenger-Charolles, & Colé, in press) was developed as in the *Word Frequency Book* by Carroll et al. (1971), and more recently, the *Educator's Word Frequency Guide* (Zeno, Ivens, Millard & Duvvuri, 1995) used in English for child language studies. It is based on a corpus of 1.9 million words extracted from 54 readers used in French primary schools, between the first and fifth grades. The readers cover a range of topic areas, each with an appreciable amount of data coming from different types of texts (from novels to various kinds of fiction, from newspaper reporting to technical writing, and from poetry to theater plays) written by different authors from a variety of backgrounds. Grammar books used in the classroom were also included in the *Manulex* corpus. The database contains two lexicons: the wordform lexicon (48,886 entries) and the lemma lexicon (23,812 entries). Each lexicon provides a grade-level-based list of words found in first-grade, second-grade, and third-tofifth grade readers (hereafter called levels G1, G2, G3-5, respectively). A fourth level (G1-5) was generated by combining all readers.

Lemmatization was achieved by à commercially distributed tagger called *Cordial Analyseur©.* The syntactic labeling system used by the analyzer is rather detailed, since it comprises 130 different labels, corresponding to the majority of the morpho-syntactic distinctions of French. A main set of 12 labels were retained for the distributed database.

To compare French readers and printed adult materials, we used the *Lexique* database (New et al., 2001) and some corpus statistics from *Cordial. Lexique is* the current reference tool in French psycholinguistic research. It is based on a corpus of 31 million words and contains 128,942 wordform entries (inflected forms of verbs, nouns, and adjectives) and 54,196 lemma entries. Proper names, symbols, abbreviations, and foreign language words are not included. The *Cordial* statistics come from a corpus of novels and essays (130 and 22 million words, respectively). These statistics are given in the software for comparison of an tagged input text to a reference corpus chosen by the user.

[1] For: ***Lex****ique des* ***Manu****els,* i.e. the lexicon of readers.

Table 1. Lemma and Wordform Counts for Open-Class and Closed-Class Labels in *Manulex*

	Lemmas					Wordforms			
	G1	G2	**G3-5**	**G1-5**	***Lexique****	**Gl**	G2	**G3-5**	**GI-5**
Open-class									
Noun	3,520	5,149	10,366	10,837	34,393	42,940	79,606	319,658	442,204
Verb	1,180	1,751	3,083	3,158	18,957	30,280	65,018	248,784	344,082
Adjective	930	1,689	4,167	4,317	18,252	9,654	24,062	99,303	133,019
Adverb	233	362	713	725	1,685	9,226	21,366	81,285	111,877
Proper name	625	1,207	3,780	4,454	-	5,942	8,356	31,681	45,979
Interjection	78	89	123	139	202	812	1,012	2,777	4,601
Closed-class									
Determiner	14	17	18	18	18	32,137	58,934	229,919	320,990
Conjunction	19	21	23	23	30	5,780	14,278	57,565	77,623
Preposition	38	44	52	53	79	15,364	34,286	151,643	201,293
Pronoun	56	57	61	61	84	19,959	43,275	161,141	224,375
Total	**6,693**	**10,386**	**22,386**	**23,785**	**73,700**	**172,094**	**350,193**	**1,383,756**	**1,906,043**

*: New et al.'s (2001) adult database. The counts were computed from the lemma lexicon, where entries are collapsed by orthographic form (54,196 entries). In *Manulex,* entries are collapsed by syntactic category. So, for comparison, *Lexique's* lemma entries were expanded as in Manulex (73,700 entries).

Table 1 provides lemma and wordform counts for open-class and closed-class items (24 abbreviations and 3 euphonic strings were omitted from this analysis). French readers, as represented in *Manulex* by materials for grades 1 to 5, contains 23,785 distinct lemmas [2]. These lemmas result in a total of nearly 1.9 million graphically distinct wordforms. The vocabulary growth between levels G1 and G2 is 3,693 lemmas, and between levels G2 and G3-5 is 12,000 lemmas. Compared to New et al.'s (2001) database, the G1-5 lemmas cover 32% of the nouns found in adult printed materials, 17% of the verbs, 24% of the adjectives and 43% of the adverbs.

The above counts give the absolute vocabulary size of a French child, that is, the words he or she has to deal with between grades 1 and 5. Nagy and Anderson's (1984) study certainly overestimated the size of the vocabulary proposed in schools to an English-speaking child. In fact, their study was based on a projection of a random sample of 7,260 words from the 86,741 words in the Carroll et al. (1971) *Word Frequency Book.* Our analysis better reflects child print exposure because it is based on a corpus of French readers used in nearly 75% of all French primary schools. Given the fact that Nagy and Anderson's estimate was based on word families, which reduces the number of items compared to our lemma count (see Footnote 2), the overestimation seems very high: our G3-5 level contains 22,386 different lemmas, whereas Nagy and Anderson's count suggests 88,500 word families in grades 3 to 9. The difference cannot come from grades 6 to 9 because *Lexique* (computed from a corpus of adult printed materials containing 31 million words) contains 73,700 lemmas (see Table 1 note) which is below their printed school English estimate.

Table 2 compares the proportion of closed-class words and open-class words in *Manulex* and *Cordial.* For both corpora, one half of the words are closed-class words, and one half are open-class words. The open-class word distributions in *Cordial* corpora and each *Manulex* level were compared using a chi-square test, but no significant values were found, indicating that the different samples could be considered as having been taken from a same corpus.

Table 2. Proportion of Closed-Class and Open-Class Words in *Cordial* and *Manulex* Corpora

	Cordial		***Manulex***			
	Novels	**Essays**	**Gl**	G2	**G3-5**	**G1-5**
Closed-class words	50.0	50.3	46.1	47.3	50.2	48.8
Open-class words	50.0	49.7	54.0	52.7	49.8	51.2
Nouns	44.7	47.3	53.9	48.9	52.1	51.8
Adjectives	11.1	13.0	8.6	10.3	11.4	10.6
Verbs	29.6	25.5	27.7	29.3	24.6	26.1
Adverbs	14.6	14.1	9.8	11.5	11.9	11.4

The number of different words French children actually encounter in readers can be also described using a reader-based analysis. In fact, a given reader better reflects what children

[2] A word family (Bauer & Nation, 1993; Nagy & Anderson, 1984) includes the base word, all of its inflections, and its common derivatives. The term lemma is more restricted and includes only the base

are exposed to during a school year. Table 3 provides the mean number of lemmas and wordforms found in each reader. Between absolute vocabulary size, given by the database, and reader-based vocabulary size, there is a drop of about one seventh (23,785 vs. 3,680 for G1-5). A child gains about 1,000 words (lemmas) in second grade and about 2,000 words in third to fifth grade. Compared to the 73,700 lemmas found in an adult corpus, the G3-5 child has the opportunity to experience only one fifteenth of these words if he or she only read the school reader.

Table 3. Mean Number of Wordforms and Lemmas in French Readers

	G1	**G2**	**G3-5**	**G1-5**
Mean number of wordforms	13,258	27,002	49,520	35,369
Range	4,554-18,507	8,979-44,600	18,918-97,173	4,554-97,173
Mean number of lemmas	1,870	2,879	4,893	3,680
Range	1,076-2,529	1,740-4,277	3,006-7,732	1,076-7,732
Mean number of wordforms per lemmas	**7**	9	10	10

Based on the number of wordforms, the mean number of encounters of each lemma is about 10. What does the research say about the number of encounters needed to learn a word? Saragi, Nation and Meister (1978) found that words presented to learners fewer than six times were learned by half of their subjects, while words presented six or more times were learned by 93%, suggesting a threshold of six encounters. In a review published a few years later, Nation (1982) noted that a figure of 16 encounters was common in the literature. In an empirical study, Jenkins, Stein and Wysocki (1984) found that only about 25% of learners had learned a word after 10 encounters. Working with a different metric, Nagy, Herman and Anderson (1985) estimated the likelihood of acquisition after one encounter to be about .15, with full acquisition occurring in six or seven encounters. In a follow-up to this experiment, however, Herman, Anderson, Pearson and Nagy (1987) lowered the estimate to .05 for authentic texts, with full acquisition in 20 encounters. A more recent study by Rott (1999) using experimental texts located the critical cutoff point at six encounters, which brings us right back to Saragi et al.'s (1978) figure. Given such conflicting measures and measurements, Zahar, Cobb and Spada (2001) consider that no studies provide definitive evidence of how many exposures to a word are needed to ensure its learning.

Our report of printed school French led to a considerably smaller estimation of vocabulary use than in the often-cited Nagy and Anderson (1984) study. Our counts lead to an average of 10 encounters for each lemma. Given that a great number of words are often repeated, children must read a lot of printed material to build their mental lexicon by incidental learning. To reach Nagy and Anderson's (1984) one million words, a French third to fifth grader would have to read the equivalent of 20 readers per year (1.5 a month). Cobb (1997) pointed out that logically, incidental learning over time makes sense. In texts, words are visible, noticeable, repeated, and so over time, a learner will come across words in every

word and its inflections. It corresponds to the basic set of syntactic categories in the *Manulex* lexicons. Word families would give less coverage.

type of context. With an exposure to an average of a million words of text per year at school, incidental acquisition can be shown by simple arithmetic to account for the size of an adult vocabulary. However, not all children read that much, even though children should be encouraged to read as much as possible.

Word Frequency Computations in MANULEX

Although the language makes use of a large number of words, not all of those words are equally useful. One measure of usefulness is word frequency, that is, how often the word occurs in a corpus. Clearly, no actual corpus employed for any given word count can contain all words in the language; however, the larger the corpus, the greater the proportion of words included in the count, and the stabler the relative frequencies of rare words.

High-frequency words are processed more quickly than low-frequency words. This is the longest-standing and most clearly established result in experimental psycholinguistics. It has been replicated in a range of tasks such as reading aloud, picture naming, and semantic or lexical decision making, and in a range of languages. The word frequency effect is taken as evidence that the systems involved in language processing respond to basic statistical properties of an individual's linguistic experience (for a review, see Monsell, 1991).

The frequency computations made for *Manulex* used the methods described in Carroll et al. (1971) and Zeno et al. (1995) (see also Breland, 1996), with four indices at each grade level: F: for the overall word frequency; *D:* for the index of dispersion across readers; *U:* for the estimated frequency per I million words, derived from F with an adjustment for *D;* and *SFI:* for the standard frequency index, derived directly from *U*. A word type with an *SF1* of 40 occurs once in a million tokens; one with *SFI* = 50 occurs 10 times per million; one with an *SFI* of 60 occurs 100 times per million, etc.

The grade-level-based frequency computations were weighted by the index of dispersion across the readers. In this way, words recurring in a single reader, can be distinguished from words recurring in many readers. This gives a better estimate of the true frequency 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. Thus, for the same overall frequency F, the dispersion index gives to an estimated frequency value U that is twice as high. Using Zeno et al.'s counts in a study on age of acquisition, Zevin and Seidenberg (2002) showed that word frequencies were more closely correlated with latencies than were earlier counts such as those extracted from adult corpora like the Kucera and Francis's (1967) norms. This suggests that for language acquisition research, precise child norms must be used to measure how often words are experienced by children during their exposure to print.

In short, the measure of frequency proposed here is based on a combination of 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. Each reader is considered as a semantically coherent piece of text and it is assumed that different readers can be interpreted as different contexts. An even better approximation of context could have been achieved by segmenting each reader into chapters.

Distribution of Frequent and Rare Words

The percentage of words (lemma count) occurring only once (F=1), a phenomenon known as *hapax legomena, is* 29% in levels G1 and G2, 24% in G3-5, and 23% in G1-5. *Hapax dislegomena (F = 2)* is 13% in G1, 15% in G2,14% in G3-5, and 13% in G1-5. Thus, about one third of what children read is composed of rare events (the same proportion is found in adult corpora). The high number of low-frequency words in child corpora show that reading is a particularly efficacious way to expand a child's vocabulary.

Cunningham and Stanovich (2001) (see also Stanovich et al., 1996) reported a study by Hayes and Ahrens (1988) which analyzed word distributions in various written and spoken contexts. The written language was sampled from genres as difficult as scientific articles and as simple as preschool books; the spoken words were taken from television shows of various types; and the adult speech was from two contexts varying in formality. Table 4 displays Hayes and Ahrens' (1988) data, along with our counts from *Manulex* and *Corpaix,* a French adult speech database (Véronis, 2003). The column labeled "Rank of Median U" is the rank of the word with the corresponding U value (weighted frequency per million) in Carroll et al.'s database (English language) and in New et al.'s database (French language). So, for Hayes and Ahrens' data, the median U in children's books corresponds to the 627`h most frequent word in the Carroll et al. count (86,741 entries); the median U in popular magazines is ranked the 1,399`h most frequent; and the median U in abstracts of scientific articles, not surprisingly, has a very low rank (4,389). For the French data, the median U in level G 1 was ranked the 4,264`" most frequent in the New et al. count (54,149 entries), and 4,756`h when level G1-5 was considered. (Note that the ranks cannot be compared between the two language data sets because the number of entries, and hence the number of ranks, are different in the two databases.)

What is immediately noticeable is how lexically impoverished speech is compared to the written word language. With the exception of the special situation of courtroom testimonies, the average frequency of all words in the two samples of English speech is quite low, hovering in the 400-600 range. Incontestably, the words used in children's books are much rarer than those in the speech of prime-time adult television. Our analysis led to a similar pattern, with a median U in adult conversation corresponding to the 2,942`h most frequent word in the New et al. database. French reading books contain words that are twice as rare as those heard in adult conversations. As pointed out by Stanovich et al. (1996), these differences in word rarity have direct implications on vocabulary development. If most vocabulary is acquired outside of formal teaching, the only opportunities to acquire new words occur when an individual is exposed to a word in the written or oral language that is outside his/her current vocabulary; *"this will happen vastly more often while reading than while talking or watching television" (p.* 20).

On the opposite side, what is the proportion of frequent events? A language has the tendency to recycle a small number of words over and over again, and as stated by Nation and Warring (1997), if children know these words then their reading power can be greatly enhanced for a relatively modest learning investment. Table 5 shows that in the *Manulex* corpora, a few lemmas account for most of the wordforms at any level. Each level was rank-ordered by U and cumulative percentages were computed. Ten words account for 30% of the ink on any page (repeated words like *"le"* and "de"), and only 1,000 lemmas account for more than 84% of the wordforms, 2,000 accounting for more than 87% (93% in G1).

Table 4. Rank of Median U (Frequency Per Million) in Selected Sources of Written and Spoken English and French

	Rank of median U
1. Printed texts	
English: Hayes & Ahrens's (1988) data	
Abstracts of scientific articles	4,389
Newspapers	1,690
Popular magazines	1,399
Adult books	1,058
Comic books	867
Children's books	627
Preschool books	578
French: *Manulex* database	
G 1	4,264
G2	4,552
G3-5	4,734
G1-5	4,756
2. **Adult speech**	
English: Hayes & Ahrens's (1988) data	
Expert witness testimonies	1,008
College graduates to friends, spouses	496
French: *Corpaix* corpus	2,942

Table 5. The Percentage of Text Coverage of Each Successive Rank-Ordered Lemma in the Three Levels of *Manulex*

	Percent of wordforms		
Rank-ordered lemmas	**G1**	G2	**G3-5**
10	30	29	28
100	57	58	56
500	76	77	72
1,000	85	84	80
2,000	93	91	87
3,000	96	94	91

Thus, a French first-grade child who knew 1,000 words (lemmas) would understand nearly 80% of the wordforms found in the G3-5 corpora. With a vocabulary size of 2,000 words, he or she would understand 87% of the wordforms in a text, which means that less than one word in every ten would be unknown.

The above rare and frequent word distributions demonstrate the quantitative linguistic universal that, in all languages, a very small number of frequent words cover the majority of a text, while a very large number of rare words only cover a tiny proportion of the text. However, as pointed out before, to know these words, the child needs to encounter them

frequently and in a variety of contexts. To build up his/her lexicon, the child should read a lot of printed material because the probability of encountering a new word is 1 in 10. Analyses also illustrate that rare and frequent words do not have the same connotation in language, and this is an important fact for understanding their impact in lexical growth. In an analysis of old English, Erilt (1999) pointed out that frequent words, in addition to having concrete meanings, tend to also have a greater number of abstract meanings than rare words. This can be partly explained by the influence of the age of the words, in such a way that the most frequent words are also the oldest ones and have, by a metaphorical process, developed abstract meanings in addition to concrete ones. The addition of abstract meanings also correlates with polytextuality, i.e., the occurrence of a word in various text types. This principle applies not only to nouns but also to verbs, which acquire new combinations of arguments in the course of time. These ideas can be developed further in the context of lexicon acquisition. Given the distribution of high-frequency words, children are exposed to more polysemous words. Teaching these words can help children distinguish meanings and enlarge their vocabulary.

An Example of a Basic Vocabulary for Teaching

A constant preoccupation of language teachers is to identify which words should be given priority in their teaching, and many researchers have suggested that teachers should first concentrate on high-frequency words. This means that they need to have reference lists to judge whether or not a particular word deserves attention and whether a text is suitable for a class. They also need some useful vocabulary lists based on frequency. Some researchers consider range, that is, the occurrence of a word across several subsections of a corpus. The most well known is Carroll et al.'s (1971) *Word Frequency Book* where the main values of the list are the frequency of each word in each grade and in each subject area.

West (1953) found that frequency and range alone were not sufficient for deciding what goes into a basic vocabulary list designed for teaching purposes. This author made use of ease or difficulty of learning (it is easier to learn another related meaning for a known word than to learn another word), necessity (words that express ideas that cannot be expressed through other words), coverage (it is not efficient to be able to express the same idea in different ways; it is more efficient to learn a word that covers a quite different idea), and stylistic level and emotional words (West saw second language learners as initially needing neutral vocabulary).

To supply the main words of French readers and to develop a basic vocabulary for French learners, perhaps the commonest criterion that we can use is the frequency of recurrence of a given lexical item across the three main levels in *Manulex* (G1 vs. G2 vs. G3-5), i.e., all overlapping entries across levels. Most word-frequency studies have acknowledge the importance of range of occurrence. However, a word should not become part of a basic vocabulary list solely because it occurs frequently: It should occur frequently across a wide range of texts. This does not mean that its frequency has to be roughly the same across the different readers, but it means that it should occur in some form or other in most of the different readers. So, to omit all context-specific words, all entries with a D value below .25 were deleted (the word was encountered in nearly one quarter of the 54 readers). Finally, we rank-ordered the list by syntactic label, including proper names. This left a total of 3,450

lemma entries. Table 6 reports the number of entries and frequency statistics for open-class and closed-class words. The Appendix gives the first 100 rank-ordered open-class words.

Table 6. Some Statistics about the Basic Vocabulary in *Manulex*

	N	**Lemma coverage** (%)				**Wordform coverage** (%)			**Mean SF1**	
		G1	**G2**	**G3-5**	***Lexique***	**G1**	**G2**	**G3-5**	**G1-5**	***Lexique***
Open-class										
Noun	1903	54.1	37.0	18.4	5.5	91.3	85.9	76.4	56	54
Verb	749	63.5	42.8	24.3	4.0	97.7	95.3	91.0	58	57
Adjective	413	44.4	24.5	9.9	2.3	91.0	83.9	72.7	56	57
Adverb	150	64.4	41.4	21.0	8.9	98.8	97.1	93.7	60	63
Proper name	87	13.9	7.2	2.3		37.3	18.1	17.5	53	
Interjection	32	41.0	36.0	26.0	15.8	71.7	76.8	80.3	53	54
Closed-class										
Determiner	13	92.9	76.5	72.2	72.2	100.0	100.0	100.0	75	78
Conjunction	19	100.0	90.5	82.6	63.3	100.0	99.9	99.8	68	69
Preposition	35	92.1	79.5	67.3	44.3	100.0	99.9	99.8	67	68
Pronoun	49	87.5	86.0	80.3	58.3	99.9	99.9	99.7	67	71
Total	3 450	51.5	33.2	15.4	4.7	94.5	92.6	88.6		

The basic vocabulary, as defined by the overlapping lemma entries in G1 through G5, covers 51.5% of G1's lemmas, 33.2% of G2's lemmas, 15.4 of G3-5's lemmas, and only 4.5% of *Lexique's* lemmas. However, the list covers 88.6% of the total wordforms found in level G3-5. The closed-class category is entirely represented in the vocabulary. This means that only one open-class word out of 10 should be unknown in a 10-year-old's reading material, and all closed-class words should be understood. The mean *SFI* points out that most items are high-frequency words (as in the adult database) with an occurrence of 10 to 100 times per million.

In short, the list accounts for at least 88% of the words on any page, whatever the grade of the reader, and this is probably the word knowledge of a French learner at the end of fifth grade. Moreover, during the five years of primary school, the 3,450 basic lemmas can easily be taught at a rate of almost 700 basic words per year.

Past Tense Frequencies in French Readers

The temporal asynchrony between grammatical and lexical development is one of the most well-established facts in developmental psycholinguistics. In every language that has been studied, children do not inflect words until they have passed through a prolonged singleword stage, where content words are used with few inflectional contrasts. For Bates and Goodman (1997), children use various morpheme discovery strategies that make use of semantic, syntactic and phonological information. It is also likely that children use the frequency and distribution of words as sources of evidence. Some statistical distributional information can be directly extracted from their linguistic contexts and can be used prior to any linguistic analysis: learners can later use more sophisticated linguistic information to

refine the coarse guesses about morphological structure made on the basis of distributional information.

The acquisition of the past tense form of English verbs has become the focus of an ongoing debate between rule-like mechanisms and a PDP single mechanism approach. In dual mechanism accounts (e.g., Marcus, 1995; Pinker, 1991, 1999), the regular past tense is produced via a rule that applies to the root stem of the verb, which is stored in the mental lexicon. Irregulars, on the other hand, are formed via associations between present and past tense forms, each of which is stored as a separate lexical entry. The existence of a separate past tense entry, as in the case of irregulars, blocks the application of the rule. However, if the representation of a past tense entry is weak (due to its low frequency), the rule may be applied erroneously, causing an over-regularization error. In addition to irregulars being regularized, regulars can also be incorrectly produced as if they were irregulars.

PDP single mechanism models (e.g., Rumelhart & McClelland, 1986) rely on a single mechanism to account for both regular and irregular past tense production. The claim is that both regulars and irregulars are represented in a single neural network. The network encodes mappings between present and past tenses as weighted links between forms. More exposure to a particular mapping strengthens the corresponding link. In this way, the mapping strength for both regular and irregular items is determined by item frequency and by the consistency of the present-to-past tense mapping within the neighborhood of phonologically similar verbs.

Table 7. **French Past Tense Frequencies in *Cordial* and *Manulex* Corpora**

		Cordial			***Manulex***		
		Novels	**Essays**	**GI**	**G2**	**G3-5**	**G1-5**
Millions of words in corpora		130.274	22.164	0.175	0.354	1.397	1.926
Indicative	Present	38.0	59.2	71.2	61.3	49.8	56.4
	Imperfect	22.1	9.6	4.2	6.5	12.6	9.7
	Past simple	17.5	8.4	4.8	7.1	12.1	9.7
	Future	2.8	3.8	2.4	3.3	3.1	3.0
	Perfect	6.7	7.2	8.3	6.9	7.1	7.3
	Pluperfect	4.1	1.7	0.3	0.7	1.7	1.2
	Past anterior	0.1	0.1	0.0	0.0	0.1	0.1
	Future anterior	0.1	0.1	0.1	0.1	0.1	0.1
Conditional	Present	2.6	3.2	1.1	1.1	1.6	1.4
	Past I	0.7	0.5	0.1	0.2	0.3	0.2
Imperative	Present	3.0	2.5	7.5	11.9	10.3	10.1
Subjunctive	Present	1.0	2.1	0.3	0.7	0.9	0.7
	Imperfect	0.7	0.8	0.0	0.0	0.1	0.1
	Past	0.2	0.4	0.0	0.0	0.1	0.1
	Pluperfect	0.4	0.4	0.0	0.0	0.1	0.0

Clearly, frequency effects contribute to understanding morphology. To illustrate, Table 7 provides past tense frequencies found in French readers as well as *Cordial* statistics. The size of each corpora is given, and for 100 conjugated verbs in the corpora, the frequencies of the past tense of the main modalities of French grammar are given. (Note that 17% of the verbs are in the infinitive in French readers.)

Chi-square values were computed to compare the indicative past-tense frequency distributions of novels and essays in *Cordial* corpora to the *Manulex* levels. The past-tense frequency distribution differs statistically between *Cordial* novel corpora and level G1 (X^2 (4) = 30.7, $p<.01$) and level G2 ($X^2(4) = 18.8$, $p<.01$), but not level G3-5 (X^2 (4) = 5.5, $p>.10$). The distributions differ by the proportion of present and imperfect tenses: there are more present and imperfect verbs in adult novels than in GI and G2 readers, where the most prevalent is the present. No significant chi-square values were found for comparisons between *Cordial* essay corpora and any *Manulex* level.

CONCLUSION

With the above corpus-based findings, it will be easy to claim that when a French learner knows 1,000 words (lemmas), hence 80% of the wordforms in a text, direct instruction in vocabulary should end; the rest can be acquired by print exposure. However, knowing 80% of the words in a text still leaves 2 out of 10 words unknown, a density probably too great for many successful inferences to take place. Indeed, researchers now believe that inferences are unlikely to be successful when only 80% of the words are known, and it only becomes consistently practical at levels more like 95% or one unknown word out of 20 (Hirsh & Nation, 1992; Laufer, 1989, 1992). Therefore, a suitable goal for primary-school children is to learn enough words to build their vocabulary.

However, the distance from 80% to 95% is larger than it seems because, to gain 15%, the learner has to progress from 1,000 words to 3,000 words (see level G3-5 in Table 5). How long does it take to learn 2,000 more words? Is it a feasible goal during one school year? Bloom (in press) states that children clearly learn at least some words in a punctual manner; but they do not need days, weeks, months, or years of cumulative experience to learn them. They clearly also learn at least some words in a prolonged fashion, building their meaning in the course of several years. Bloom gives the examples of verbs such as *"pour"* and *"fill"*, which undergo a semantic development process that lasts until adulthood, just like words such as *"game"*, *"democracy"*, and so on. In sum, even when one is careful to allow for changes in the word-learning rate through vocabulary instruction, the idea that word learning or teaching can be accurately summed up as *"x-year-olds learn y words per day" is,* for Bloom (in press), overly simple and provides a false impression of how words are learned.

What about the hypothesis of incidental learning by print exposure? Our corpus-based description of the lexicon addressed to a child is inconsistent with Nagy and Anderson's (1984) often-cited study. In a reader-based count (Table 3), we showed that a French child is exposed to about 2,000 word lemmas in G1, 3,000 in G2, and 5,000 in G3-5. In an absolute count *(Manulex* lemma count, Table 1), we showed that a French child can be exposed to about 7,000 word lemmas in G1, 10,000 in G2, and 22,000 in G3-5. We are far from Nagy and Anderson's estimate of about 88,000 word families. Moreover, the 3,000 lemma increase between G1 and G2 have a mean U(weighted frequency per million) of 2.81, with a median

of .43; and the 12,000 lemmas increase between G2 and G3-4 have a mean U of .87, with a median of .09. (The mean U of the G1 lexicon is 123.42 and the median is 5.56.) So, the child is exposed to more words during their grade-school years but it is a set of extremely rare words, occurring essentially once in a reader.

Huchin and Coady (1999) reviewed the empirical research on the issue of incidental vocabulary acquisition. They claim that much depends on the context surrounding each word, the nature of the learner's attention, and task demands, which affect incidental vocabulary acquisition. Gass (1999) goes further and argues that because the definitions of "incidental", "vocabulary", and "acquisition" are unclear, it is difficult to support the notion of incidental vocabulary leaming.

As pointed out by Rack, Hulme and Snowling (1993), we still do not know the answers to basic questions such as how vocabulary size typically changes over time and what mechanisms and/or skills foster its growth at different points in time. There is no doubt that lexical representations can be derived from statistical distribution cues through exposure to print, but there is also no doubt that acquiring lexical knowledge requires a lot of effort for the child, particularly time spent in reading to connect new lexical information to background knowledge. As a consequence, helping children learn about words is certainly the best way to build their lexicon.

However, the statistical learning framework and the PDP view remain valuable theoretical advancements that help us understand vocabulary development. What PDP networks have changed in our conception of the mental lexicon in particular, and cognitive processing in general, is that there is no need to represent lexical knowledge in a passive fashion as in a storeroom. In PDP, there is no lexicon and no access. Lexical knowledge is a by-product of processing; it emerges from the interconnection strengths that allow word concepts to be recreated.

Chomsky (1975) stated that *"every child comes to know facts about the language for which there is no decisive evidence from the environment. In some cases, no evidence at all."* Our corpus-based description of the lexicon to which a child is exposed shows that important statistical properties of the written word are learnable, in contradiction to his most celebrated argument, the poverty of the stimulus. However, we have outlined the idea here that the mental lexicon cannot develop without reading and without instructional actions to enhance individual word knowledge.

Clearly, statistical learning is not the only relevant mechanism for building the lexicon. Hence, the present findings should not be viewed as decisive proof that the connectionist approach is right and that the rule-based dual mechanism account is wrong. But the position I would like to take here is similar to Long and Almor's (2000) who argued: *"Rather, the present findings should be viewed as adding one piece to a growing body of evidence that suggests that the separation of language processing into two mechanisms is buying less and less in terms of explanatory power but costing more and more in terms of unnecessary theoretical baggage."*

REFERENCES

Aitchison, J. (1994). *Words in the mind.- An introduction to the mental lexicon.* (2nd Ed.) Oxford: Blackwell.

Batchelder, E.O. (1997). *Computational evidence for the use of frequency information in discovery of the infant's first lexicon.* Unpublished PhD Thesis, City University of New York, New York. Retrieved December 12, 2002, from http://www.human.tsukuba.ac.jp/ - eleanorb/diss.html#download

Bates, E., & Goodman, J.C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes, 12,* 507-584.

Bauer, L, & Nation, P. (1993). Word families. *International Journal ofLexicography, 6,* 253-279.

Beck, LL., & McKeown, M.G. (1991). Conditions of vocabulary acquisition. In P. D. Pearson (Ed.), *The handbook of reading research (Vol.* 2, pp. 789-814). New York: Longman.

Bloom, P. (in press). Myths of word learning. In D.G. Hall & S.R. Waxman (Eds.), *Weaving a lexicon.* Cambridge, MA: MIT Press.

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

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33,* 111-153.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan & G.A. Miller (Eds.), *Linguistic theory and psychological reality (pp.* 264-293). Massachusetts: MIT Press.

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

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass: MIT Press.

Chomsky, N. (1975). *Reflections on language.* New York: Pantheon.

Christiansen, M.H., Allen, J., & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221-268.

Cipielewski, J., & Stanovich, K.E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology, 54,* 74-89.

Clark, E.V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T.E. Moore (Ed.), *Cognitive development and the acquisition of language.* New York: Academic Press.

Clark, E.V. (1993). *The Lexicon in Acquisition.* Cambridge: Cambridge University Press.

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.

Cobb, T. (1997). *From Concord to lexicon: Development and test of a corpus-based lexical tutor.* Unpublished PhD Thesis, Department of Educational Technology, Concordia University, Montreal, Quebec. Retrieved April 14, 2002, from http://www.er.uqam.ca/ nobel/r21270/webthesis/Thesis0.htm1

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.

Cunningham, A.E., & Stanovich, K.E. (1991). Tracking the unique effects of print exposure in children: Associations with vocabulary, general knowledge, and spelling. *Journal of Educational Psychology, 83,* 264-274.

Cunningham, A.E., & Stanovich, K.E. (1993). Children's literacy environments and early word recognition skills. *Reading and Writing: An Interdisciplinary Journal, 5,* 193-204.

Cunningham, A.E., & Stanovich, K.E. (2001). What reading does for the mind. *Journal of Direct Instruction, 2,* 137-149.

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

Elman, J.L. (1995). Language as a dynamical system. In R.F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition (pp.* 195-223). Cambridge, MA: MIT Press.

Erilt, L. (1999). *The frequency structure of language with special reference to old English.* Unpublished MA Thesis, Department of English, University of Tartu. Retrieved June 13, 2002, from http://www.ehi.ee%lumme/mag/

Forster, K. (1976). Accessing the mental lexicon. In R.J. Wales & E. Walker (Eds.), *New approaches to language mechanisms (pp.* 257-287). Amsterdam: North-Holland.

Gambrell, L.B. (1996). Creating classroom cultures that foster reading motivation. *The Reading Teacher, 50,* 14-25.

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition, 21,* 319-333.

Grainger, J., O'Regan, J.K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics, 45,* 189-195.

Grainger, J., O'Regan, J.K., Jacobs, A.M., & Segui, J. (1992). Neighborhood frequency effects and letter visibility in visual word recognition. *Perception & Psychophysics, 51,* 49-56.

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

Hayes, D.P., & Ahrens, M.G. (1988). Vocabulary simplification for children: A case of motherese. *Journal of Child Language, 15,* 395-410.

Hebb, D.O. (1949). *The organization of behavior. A neuropsychological theory.* New York: John Wiley.

Herman, P.A., Anderson, R.C., Pearson, P.D., & Nagy, W.E. (1987). Incidental acquisition of word meaning from expositions with varied text features. *Reading Research Quarterly, 22,* 263-284.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8,* 689-696.

Huchin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition, 21,* 181-193.

Hulme, C., Snowling, M.J., & Quinlan, P. (1991). Connectionism and learning to read: Steps towards a psychologically plausible model. *Reading and Writing, 3,* 159-168.

Jerkins, J.R., Stein, M.L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21,* 767-787.

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

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking (pp.* 316323). Clevedon: Machines Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics (pp.* 126-132). New York: Macmillan.

Lété, B., Sprenger-Charolles, L., & Colé, P. Manulex (in press). A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers.*

Long, C., & Almor, A. (2000). Irregularization: The interaction of frequency and phonological interference in regular past-tense production. *Proceedings of the twentysecond annual conference of the cognitive science society (pp.* 310-315). University of Pennsylvania. Retrieved January 6, 2003, from http://www-scf.usc.edu/-clong/Long. Almor.CogSci2000 fma1.PDF

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203-208.

Lund, K., Burgess, C., & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the Cognitive Science Society (pp.* 660-665). Hillsdale, N.J.: Erlbaum.

Marcus, G.F. (1995). The acquisition of the English past tense in children in multilayered connectionist networks. *Cognition, 56,* 271-279.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of contexts effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 365-407.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53,* 253-279.

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.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76,* 165-178.

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

Nagy, W.E., & Herman, P.A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M.G. McKeown & M.E. Curtis (Eds.), *The nature of vocabulary acquisition.* Hillsdale, NJ: Erlbaum.

Nagy, W.E., Herman, P., & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly, 20,* 233-253.

Nation, P. (1982). Beginning to learn foreign vocabulary: A review of the research. *RELC Journal, 13,* 14-36.

Nation, P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle & Heinle Publishers.

Nation, P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing, 10,* 27-40.

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

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy (pp.* 6-19). New York: Cambridge University Press.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur Internet: Lexique. *L'Année Psychologique, 101,* 447-462.

Oldfield, R.C. (1966). Things, words and the brain. *Quarterly Journal of Experimental Psychology, 18,* 340-253.

Paap, K.R., Newsome, S.L., McDonald, J.E., & Schvaneveldt, R.W. (1982). An activation-verification model for letter and word recognition. *Psychological Review, 89,* 573-594.

Parry, K., (1991). Building vocabulary through academic reading. *TESOL Quarterly, 25, 629*653.

Pinker, S. (1991). Rules of language. *Science, 253,* 530-535.

Pinker, S. (1994). *The language instinct.* New York: HarperCollins. Pinker, S. (1999). *Words and rules.* New York, NY: Basic Books.

Plaut, D.C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language, 52,* 25-82.

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.

Rack, J, Hulme, C., & Snowling, M. (1993). Learning to read: A theoretical synthesis. In H. Reese (Ed.), *Advances in child development and behavior (Vol.* 24, pp. 99-132). New York: Academic Press.

Rayner, K., Foorman, B.R., Perfetti, E., Pesetsky, D., & Seidenberg, M.S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2,* 31-74.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22,* 425-469.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21,* 589-619.

Rumelhart, D.E., & McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60-94.

Rumelhart, D.E., & McClelland, J.L. (1986). On the learning of past tense of English verbs. Implicit rules or parallel distributed processing? In J.L. McClelland, D.E. Rumelhart and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Saragi, T., Nation, P, & Meister, G.F. (1978). Vocabulary learning and reading. *System, 6, 72*78.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge, UK: Cambridge University Press.

Schmitt, N., & McCarthy, M. (Eds.) (1997). *Vocabulary: Description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

Seidenberg, M.S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 1599-1603.

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

Seidenberg, M.S., McDonald, M.C., & Saffran, J.R. (2002). Does grammar start when statistics stop? *Science, 298,* 553-554.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

Stanovich, K.E. (1993). Does reading make you smarter? Literacy and the development of verbal intelligence. In H. Reese (Ed.), *Advances in child development and behavior (Vol.* 24, pp. 133-180). San Diego, CA: Academic Press.

Stanovich, K.E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford.

Stanovich, K.E., & Cunningham, A.E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20, 51*68.

Stanovich, K.E., & Cunningham, A.E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *Journal of Educational Psychology, 85,* 211-229.

Stanovich, K.E., & West, R.F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24,* 402-433.

Stanovich, K.E., West, R.F., Cunningham, A.E., Cipielewski, J., & Siddiqui, S. (1996). The role of inadequate print exposure as a determinant of reading comprehension problems. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention (pp.* 15-32). Mahwah, NJ: Erlbaum.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R.J. (1987). Most vocabulary is learned from context. In M.G. McKeown & M.E. Curtis (Eds.), *The nature of vocabulary acquisition.* Hillsdale, NJ: Erlbaum.

Véronis, J. (2003). *Speech word frequencies in Corpaix.* Retrieved January 12, 2003, from http://www.up.univ-mrs.fr/weronis/

Walberg, H., & Tsai, S. (1983). Matthew effects in education. *American Educational Research Journal, 20,* 359-373.

Waring, R. (2003). *Connectionism and second language vocabulary.* Retrieved December 11, 2002, from http://wwwl.harenet.ne.jp/-waring/papers/connect.html

West, M. (1953). Is a textbook really necessary? *ELTJournal, 8,* 64-67.

West, R.F., & Stanovich, K.E. (1991). The incidental acquisition of information from reading. *Psychological Science, 2,* 325-330.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review, 57,* 541-572.

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 ofMemory and Language, 47, 1-29.*

AUTHOR'S NOTE

Partial support for this research was provided by a grant from the Ministère de la Recherche "Ecole et Sciences Cognitives".

APPENDIX

First 100 Rank-Ordered Open-Class Words In *Manulex's* Basic Vocabulary.

Rank	Nouns	Verbs	Adjectives	Adverbs	Proper names	Interjections
1	mot	être	petit	ne	paris	ah
2	phrase	avoir	tout	pas	pierre	oh
3	nom	faire	deux	plus	france	eh
4	jour	dire	quel	bien	noël	attention
5	enfant	aller	grand	très	sophie	bon
6	texte	pouvoir	bon	si	jean	tiens
7	temps	voir	chaque	tout	julien	hélas
8	maison	trouver	autre	alors	nicolas	merci
9	eau	mettre	trois	là	julie	comment
10	histoire	vouloir	beau	aussi	arthur	hé
11	fois	prendre	quelque	peu	paul	hop
12	verbe	venir	gros	encore	afrique	allô
13	chat	écrire	même	non	françois	bravo
14	chien	donner	premier	même	michel	ha
15	homme	lire	seul	toujours	thomas	ô
16	heure	savoir	suivant	vite	amérique	ouf
17	tête	falloir	blanc	jamais	soleil	chut
18	maman	regarder	vrai	oui	marie	gare
19	animal	passer	vieux	pourquoi	olivier	adieu
20	nuit	arriver	rouge	beaucoup	victor	vive
21	main	devoir	vert	trop	georges	aïe
22	terre	jouer	noir	comment	marion	là
23	yeux	manger	quatre	est-ce que	martin	salut
24	question	partir	long	maintenant	europe	zut
25	oiseau	demander	nouveau	comme	alain	plouf
26	ami	entendre	jeune	ici	alice	ça
27	chose	répondre	dernier	déjà	marc	clac
28	lettre	appeler	bleu	ainsi	guillaume	ohé
29	point	parler	cinq	souvent	italie	chouette
30	coup	chercher	joli	loin	chine	chic
31	air	aimer	plusieurs	enfin	henri	pardi
32	monsieur	sortir	plein	autour	eiffel	bof
33	arbre	tomber	possible	moins	éric	
34	monde	rester	certain	mieux	marseille	

Rank	Nouns	Verbs	Adjectives	Adverbs	Proper names	Interjections
35	exemple	retrouver	sûr	en	valérie	
36	père	penser	différent	fort	alexandre	
37	loup	crier	pauvre	peut-être	asie	
38	porte	comprendre	heureux	combien	pauline	
39	fille	ouvrir	dix	longtemps	alpes	
40	pied	courir	content	aujourd'hui	charles	
41	livre	laisser	chaud	bientôt	espagne	
42	papa	commencer	haut	vraiment	nice	
43	peur	arrêter	simple	assez	nathalie	
44	mer	attendre	six	tard	lyon	
45	place	devenir	mauvais	tant	rose	
46	groupe	suivre	passé	mal	jeannot	
47	matin	porter	énorme	puis	jérôme	
48	vent	vivre	froid	ensuite	poucet	
49	personne	croire	doux	d'abord	isabelle	
50	soleil	revenir	faux	parfois	andré	
51	soir	entrer	neuf	soudain	caroline	
52	maître	servir	gris	presque	sylvie	
55	garçon	tenir	fort	pourtant	hélène	
56	feu	perdre	complet	seulement	pyrénées	
57	bois	raconter	léger	demain	benjamin	
58	cheval	monter	drôle	surtout	nantes	
59	an	rire	nombreux	doucement	méditerranée	
60	tour	marcher	jaune	lentement	cécile	
61	ville	poser	difficile	au-dessus	lili	
62	fleur	choisir	court	rapidement	bretagne	
63	voiture	lever	aucun	tellement	rené	
64	exercice	rendre	triste	après	toulouse	
65	petit	dormir	sept	plutôt	denis	
66	moment	tirer	cent	devant	gérard	
67	couleur	utiliser	rapide	juste	béatrice	
68	mère	pousser	magique	haut	élodie	
69	fin	aider	frais	partout	cendrillon	
70	forêt	indiquer	malade	avant	versailles	
71	voix	écouter	fin	avec	émilie	
72	réponse	finir	clair	hier	zorro	
73	sens	sentir	bas	près	tortue	
74	dessin	recopier	dangereux	derrière	léon	
75	bruit	remplacer	important	debout	loïc	
76	ordre	essayer	cher	autant	stéphane	
77	page	cacher	fou	ailleurs	clara	
78	jardin	rentrer	juste	autrefois	christian	
79	feuille	sauter	rond	chaud	joël	
80	bout	tourner	mille	dehors	donald	

Rank	Nouns	Verbs	Adjectives	Adverbs	Proper names	Interjections
81	famille	observer	terrible	tôt		
82	côté	préparer	droit	là-bas		
83	rue	terminer	facile	cependant		
84	poisson	expliquer	sauvage	environ		
85	patte	asseoir	lourd	également		
86	ciel	relever	prêt	depuis		
87	roi	acheter	étrange	sûrement		
88	route	décider	un	heureusement		
89	suite	chanter	rose	dessus		
90	partie	montrer	huit	simplement		
91	frère	apprendre	sec	afin de		
92	bateau	oublier	tel	brusquement		
93	parent	dessiner	méchant	attentivement		
94	vie	voler	épais	exactement		
95	train	apercevoir	deuxième	tranquillement		
96	lit	jeter	immense	complètement		
97	table	placer	extraordinaire	là-haut		
98	pays	boire	magnifique	point		
99	neige	descendre	joyeux	finalement		
100	gens	approcher	bizarre	quelquefois		