Behavior Research Methods, Instruments, & Computers
1999, 31 (2), 376-379

LEXOP: A lexical database providing orthography–phonology statistics for French monosyllabic words

RONALD PEEREMAN
Université de Bourgogne and CNRS, Dijon, France

and

ALAIN CONTENT
Université Libre de Bruxelles, Brussels, Belgium

During the last 20 years, psycholinguistic research has identified many variables that influence reading and spelling processes. We describe a new computerized lexical database, LEXOP, which provides quantitative descriptors about the relations between orthography and phonology for French monosyllabic words. Three main classes of variables are considered: consistency of print-to-sound and sound-to-print associations, frequency of orthography–phonology correspondences, and word neighborhood characteristics.

Advances in psycholinguistic research have been accompanied by an increasing demand for lexical databases and statistical descriptions of language properties. Their most obvious contribution belongs to stimulus selection and control in empirical studies. In addition, quantitative analyses of lexical characteristics deliver fine-grained measures of theoretically relevant variables, thus potentially enhancing the power and precision of current accounts. Lexical databases also constitute valuable tools for exploring language structure through the assessment and quantification of regularities and covariation patterns (see, e.g., Berndt, Reggia, & Mitchum, 1987; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Frauenfelder, Content, Goldman, & Meunier, 1995; Kessler & Treiman, 1997; Stanback, 1992; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995).

The aim of the present paper is to describe a new computerized lexical database, called LEXOP, which provides a detailed characterization of French monosyllabic words in terms of the relations between their written and their spoken characteristics. LEXOP was developed conjointly at the Free University of Brussels and the University of Bourgogne as a research tool to facilitate stimulus selection in various experimental projects concerning reading and spelling processes. Our effort was motivated by the relative scarcity of objective, statistical descriptions of the mapping between phonology and orthography for the French language (but see Véronis, 1986; Ziegler, Jacobs, & Stone, 1996). Moreover, whereas each of the available studies focuses on a single variable, our objective was to provide a more exhaustive description of relevant characteristics, estimated from the same word corpus.

In particular, one interesting aspect of the present project is that all the computations were performed not only on print-to-sound but also on sound-to-print relations. Obviously, estimates of sound-to-print complexity are directly relevant for studies of written production (see, e.g., Alegria & Mousty, 1994; Kreiner & Gough, 1990). Moreover, it has been suggested recently that both print-to-sound complexity and sound-to-print complexity influence reading performance (Stone, Vanhoy, & Van Orden, 1997).

The relatively broad scope of our attempt seems particularly important, given the extent of covariation between numerous psycholinguistic variables. Thus, typically, researchers selecting materials along one specific dimension would need to control for potentially confounded factors of theoretical importance.

The LEXOP database should constitute a valuable tool for psycholinguistic research on the French language, as well as for cross-linguistic investigations. First, the statistical characterization of the word corpus includes standard variables that are known to influence performance and need to be taken into account. Second, LEXOP provides information on new descriptors that have only recently been acknowledged as important, such as the different types of orthographic neighbors (Peereman & Content, 1997) or print-to-sound consistency for initial consonant (Treiman et al., 1995) and vowel (Berent & Perfetti, 1995).

This work was supported by the French CNRS, by a grant from the Direction générale de la Recherche scientifique—Communauté française de Belgique (A.R.C. Grant 96/01-203), and by a joint French– Belgian grant for scientific exchanges (Programme Tournesol). Correspondence concerning this article should be addressed to R. Peereman, Laboratoire d'Étude des Apprentissages et du Développement, CNRS, ESA 5022, Université de Bourgogne, 6, Bd Gabriel, F-21000 Dijon, France (e-mail: peereman@u-bourgogne.fr), or to A. Content, Laboratoire de Psychologie Expérimentale, Université Libre de Bruxelles, CP 191, Ave F. D. Roosevelt, 50, B-1050 Brussels, Belgium (e-mail: acontent@ulb.ac.be).

To provide information about the content of the database to the reader, we start with a short description of the word corpus. We then present a brief overview of the variables included in LEXOP and the relevant psycholinguistic literature that motivated their inclusion.

LEXICAL CORPUS

The corpus closely approximates the average monosyllabic vocabulary of speakers of French. LEXOP contains all monosyllabic word forms ($N = 2,449$) extracted from BRULEX, a computerized psycholinguistic database ($N = 35,746$) for French (Content, Mousty, & Radeau, 1990) including the word entries of the *Micro-Robert* dictionary (Robert, 1986). In addition to the 1,969 words coded as monosyllabic in BRULEX, we incorporated all words coded as bisyllabic that end in a consonant cluster + schwa (e.g. *porte*), because they may be considered as monosyllables, at least when the phonetic realization does not include a full vowel in final position (Warnant, 1987). Note that masculine and feminine forms correspond to separate entries and that homographs are distinguished only when they are nonhomophonic.

The phonological representations, extracted from BRULEX, correspond to the codes specified in the *Petit Robert* dictionary (Robert, 1987). The only change concerned the removal of the distinction between the anterior and the posterior vowels [a] and [ɑ]. This modification was motivated by the fact that the distinction is nearly completely lost in most current French dialects (Léon, 1992; Warnant, 1987). The phonological transcription is based on 15 vowels, 3 semivowels, and 19 consonants. Orthographic and phonological entries were parsed into onset (C_1), vowel (V), and coda (C_2) on the basis of phonological principles only.[1] More details about the composition of the word corpus and the segmentation algorithms appear in the manual accompanying the database.

LEXOP VARIABLES

The LEXOP database details the characteristics of the relations between orthography and phonology along three classes of variables. Two different counts were performed for each variable. In *type* counts, the values were estimated by reference to the number of relevant words in LEXOP, whereas *token* counts were weighted by the frequency of the words. Word frequency estimates were taken from the Trésor de la Langue Française norms for the second half of the 20th century (Imbs, 1971) and converted in number of occurrences per million.

A first class of variables concerns the *consistency* of the mapping between orthography and phonology for several kinds of units. Many studies have shown that print-to-sound consistency influences reading performance (see Berent & Perfetti, 1995, for a review of the English data; Content, 1991; Content & Peereman, 1992; Peereman, 1995, for French). The notion of consistency refers to the variability of the phonological codes that can be assigned to a particular orthographic unit. For example, the English vowel *oa* has different pronunciations in *road* and *broad*, and the unit *eaf* is pronounced differently in the words *deaf* and *leaf*. In general, the degree of consistency of a correspondence is estimated as the proportion of words in which the orthographic unit occurs with a particular pronunciation, relative to the total number of words, including the orthographic unit (whatever its pronunciation).

Most studies assess word consistency by reference to the *body* unit—that is, the orthographic unit corresponding to the rime and composed of the vowel and the final consonant or consonant cluster (e.g., *-eaf* in *deaf*, *-ave* in *wave*, *-ook* in *book*). However, the consistency can be analyzed for units at different levels of word structure, ranging from individual letters to the whole morpheme. In LEXOP, consistency scores were estimated for all possible units of segmentation, including onset (C_1), vowel (V), coda (C_2), C_1V (hereafter referred to as the *lead* unit; cf. Peereman & Content, 1997), and VC_2. Indeed, despite the major contribution of body–rime consistency to the naming of English words, consistency effects for the onset units and the lead units have also been reported (Kay, 1985; Taraban & McClelland, 1987; Treiman et al., 1995).

Consistency statistics in LEXOP were performed separately for orthographic-to-phonological mappings and phonological-to-orthographic mappings on C_1, V, C_2, C_1V, and VC_2 units. In addition, to enable users to select words containing highly irregular correspondences, the consistency score for the least consistent grapheme–phoneme and phoneme–grapheme correspondences in each word is also recorded.

Although far less documented than consistency effects, the *frequency of the correspondences* between orthography and phonology may also affect phonological conversion processes, and it constitutes the second series of variables. Frequency of the correspondences is merely the number of times a particular association occurs. Hence, contrary to consistency, frequency does not take into account the alternative pronunciations of the orthographic unit. Brown (1987; also, Brown & Watson, 1994) observed that body-consistent words were pronounced faster when they include a frequent body–rime correspondence. Similarly, Treiman, Goswami, and Bruck (1990) found better naming performance for pseudowords consisting of frequent body–rime correspondences. The frequency of correspondences was computed for each lexical entry in the LEXOP database and for each segmentation level in the consistency analysis. In addition, for each word, the frequency of the least frequent grapheme–phoneme (and phoneme–grapheme) correspondence is recorded. The latter variables were included inasmuch as Rosson (1985) had shown that naming performance for low-frequency words and pseudowords is affected by the frequency of the least frequent grapheme.

The third class of variables belongs to the lexical *neighborhood* of words. It has become increasingly clear over

the last few years that naming performance is influenced by the number of words orthographically similar to the target (Andrews, 1989, 1992; Peereman & Content, 1995, 1997; Sears, Hino, & Lupker, 1995). Typically, low-frequency words and pseudowords are pronounced faster when they are orthographically similar to numerous words. Orthographic neighbors are operationally defined as all the words that can be generated from the base letter string by a single letter substitution. For example, *rack*, *face*, *rice*, and *rate* are orthographic neighbors of the word *race*. This definition can be transposed to *phonological* forms. Hence, *phonological* neighbors are words obtained by a single phoneme substitution.

In a recent study (Peereman & Content, 1997), the facilitatory effect of neighborhood size was found to be determined by a subset of the orthographic neighborhood, which we called the *phonographic neighborhood*. The phonographic neighborhood is the set of words that are simultaneously orthographic and phonological neighbors of the target (e.g., *face* and *rate*, but not *rack*, are phonographic neighbors of *race*). Moreover, when partitioning the phonographic neighborhood into neighbors sharing the body (e.g., *face*), the lead (e.g., *rate*), or the consonantal skeleton (e.g., *rice*) with the target letter string (*race*), only phonographic neighbors sharing the body–rime correspondence seemed to facilitate naming. Hence, it appears from our studies that the more detailed neighborhood estimations may provide more proper control variables, although different neighborhood dimensions are obviously highly correlated (see the LEXOP user's manual for more details on the intercorrelations). The LEXOP database provides type and token counts of the size of each of these neighbor sets (orthographic, phonological, phonographic, body, lead, and consonant neighborhoods).

Finally, to facilitate stimulus selection for empirical studies, LEXOP also includes information on printed word frequency (from Imbs, 1971) and syntactic class (part of speech), as indexed in the French *Petit Robert* dictionary (Robert, 1986).[2]

AVAILABILITY

The LEXOP database and the users manual in postscript format (LexopMan.ps) can be downloaded by anonymous file transfer (ftp://ftp.ulb.ac.be/pub/packages/psyling). Frequency and consistency tables for the different units of analysis are provided in the same package. All documents are raw TEXT/ASCII files, which can be used with word processing, spreadsheet, or database softwares. Three versions are available. The Macintosh version (lexop.hqx) takes advantage of the standard fonts of the Mac OS (Geneva) for coding phonetic symbols. Transcoded versions based on the 7-bit ASCII code are also available for either PC (lexop.zip) or Unix environnements (lexop.tar.gz). Investigators using LEXOP and related data files for their research are requested to cite the present paper in their publications. Users are welcome to send comments and remarks to either author.

REFERENCES

ALEGRIA, J., & MOUSTY, P. (1994). On the development of lexical and non-lexical spelling procedures in French-speaking normal and disabled children. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling: Theory, process & intervention* (pp. 211-226). Chichester, U.K.: Wiley.

ANDREWS, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 802-814.

ANDREWS, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 234-254.

BERENT, I., & PERFETTI, C. A. (1995). A rose is a REEZ: The two-cycles model of phonology assembly in reading English. *Psychological Review*, **102**, 146-184.

BERNDT, R. S., REGGIA, J. A., & MITCHUM, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers*, **19**, 1-9.

BROWN, G. D. A. (1987). Resolving inconsistency: A computational model of word naming. *Journal of Memory & Language*, **26**, 1-23.

BROWN, G. D. A., & WATSON, F. L. (1994). Spelling-to-sound effects in single-word reading. *British Journal of Psychology*, **85**, 181-202.

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 [Brulex: A computerized lexical database for written and spoken French]. *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.), *Analytical approaches to human cognition* (pp. 213-236). Amsterdam: Elsevier.

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.

FRAUENFELDER, U. H., CONTENT, A., GOLDMAN, J.-P., & MEUNIER, C. (1995, March). *Comparative sublexical statistics: The processing units debate*. Paper presented at the 8th annual CUNY Conference, Tucson, AZ.

IMBS, P. (1971). *Études statistiques sur le vocabulaire français: Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles* [Statistical studies of the French vocabulary: Frequency table of the literary vocabulary of the 19th and 20th centuries]. Paris: Librairie Marcel Didier.

KAY, J. (1985). Mechanisms of oral reading: A critical appraisal of cognitive models. In A. W. Ellis (Ed.), *Progress in the psychology of language* (Vol. 2, pp. 73-105). London: Erlbaum.

KAYE, J., & LOWENSTAMM, J. (1984). De la syllabicité [On syllabicity]. In F. Dell, D. Hirst, & J. R. Vergnaud (Eds.), *Forme sonore du langage: Structure des représentations en phonologie* [Sound patterns in language: Structure of phonological representations] (pp. 123-159). Paris: Hermann.

KESSLER, B., & TREIMAN, R. (1997). Syllable structure and phoneme distribution. *Journal of Memory & Language*, **37**, 295-311.

KREINER, D. S., & GOUGH, P. B. (1990). Two ideas about spelling: Rules and word-specific memory. *Journal of Memory & Language*, **29**, 103-118.

LÉON, P. (1992). *Phonétisme et prononciations du français* [The phonetics and pronunciations of French]. Paris: Nathan.

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**, 382-410.

ROBERT, P. (1986). *Micro-Robert: Dictionnaire du français primordial* [Micro Robert: A dictionary of basic French]. Paris: Dictionnaires Le Robert.

ROBERT, P. (1987). *Le Petit Robert: Dictionnaire alphabétique et analogique de la langue française* [Le petit Robert: Alphabetical and analogical dictionary of the French language]. Paris: Dictionnaires Le Robert.

ROSSON, M. B. (1985). The interaction of pronunciation rules and lexical representations in reading aloud. *Memory & Cognition*, **13**, 90-99.

SEARS, C. R., HINO, Y., & LUPKER, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 876-900.

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.

STONE, G. O., VANHOY, M., & VAN ORDEN, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory & Language*, **36**, 337-359.

TARABAN, R., & MCCLELLAND, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory & Language*, **26**, 608-631.

TREIMAN, R., GOSWAMI, U., & BRUCK, M. (1990). Not all nonwords are alike: Implications for reading development and theory. *Memory & Cognition*, **18**, 559-567.

TREIMAN, R., MULLENNIX, 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.

VÉRONIS, J. (1986). Étude quantitative sur le système graphique et phono-graphique du français [A quantitative study of the orthographic and phonological system of French]. *Cahiers de Psychologie Cognitive*, **6**, 501-531.

WARNANT, L. (1987). *Dictionnaire de la prononciation française* [Dictionary of French pronunciation]. Paris: Duculot.

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

NOTES

1. The strict phonological segmentation principle adopted in parsing orthographic strings leads us to consider the letter U as part of the onset when preceded by G or Q and followed by another vowel (e.g., *gu/ide*, *qu/itte*).The parsing procedure also followed standard phonological analyses of French, in which semivowels are generally considered to be consonants (see, e.g., Kaye & Lowenstamm, 1984). However, in a few cases (n = 79), the semivowel could not be distinguished orthographically from the vowel (e.g., *oi* and *oin* in the words *froid* and *point* are pronounced /wa/ and /wɛ̃/), and the semivowel was therefore considered to be part of the vocalic unit. The same exceptional parsing also applied to two words with a prevocalic semivowel (*poele*, *moelle*) and four words with a postvocalic semivowel (*drive*, *dry*, *mile*, and *paye*).

2. Because the LEXOP and BRULEX databases (Content et al., 1990) use identical codings for orthographic and phonological representations, they can be exploited simultaneously. This is particularly interesting inasmuch as BRULEX includes additional relevant information on other psycholinguistic variables.

(Manuscript received April 15, 1997;
revision accepted for publication February 2, 1998.)