Last version : October 2, 1999

LEXOP (ver.2)

USER'S MANUAL

Ronald Peereman* & Alain Content**

Lexop is a computerized lexical database which provides quantitative descriptors of the relation between orthography and phonology for French monosyllabic words (Peereman & Content, 1999). It 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 on reading and spelling processes. 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 which have only recently been acknowledged as important, such as the different types of orthographic neighbors, or print-to-sound consistency for initial consonant and vowel. Furthermore, to attain a more general understanding of the relations between orthography and phonology, all computations were also performed on the sound-to-print mappings. The first part of the manual presents a brief description of the variables included in the database. We then describe the corpus and the computations performed. Finally, we present univariate and multivariate descriptive statistics on the set of lexical variables existing in LEXOP.

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 mappings between orthography and phonology. The notion of consistency pertains to the multiplicity of 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. The proportion of words in which a given unit occurs with a given pronunciation, relative to the total number of words including that particular unit provides an estimate of the degree of consistency of a correspondence.

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.

* Laboratoire d'Etude des Apprentissages et du Développement, C.N.R.S., ESA 5022, Université de Bourgogne, 6 bd Gabriel, Dijon, France. Email: peereman@u-bourgogne.fr.

** Laboratoire de Psychologie Expérimentale, Université libre de Bruxelles, Ave. F.D. Roosevelt, 50, B-1050 Brussels, Belgium. Email: acontent@ulb.ac.be.

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 permit the selection of words containing highly irregular correspondences, the consistency of the least consistent grapheme-phoneme correspondence and phoneme-grapheme correspondence in each word is recorded.

Although far less documented than consistency effects, the *frequency of the correspondences* between orthography and phonology may also affect phonological computation, and 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 the alternative pronunciations of the orthographic unit into account. The frequency of correspondences was computed for each lexical entry in the LEXOP database and for each segmentation level as for the consistency analysis. In addition, for each word, the frequency of the least frequent grapheme-phoneme (and phoneme-grapheme) correspondence is recorded.

The third class of variables belongs to the lexical *neighborhood* of 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, termed the phonographic neighbors, which are also 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. The LEXOP database provides counts of the size of each of these neighbor sets.

To facilitate stimulus selection for empirical studies, LEXOP also includes information on printed word frequency (from Imbs, 1971), and syntactic class as indexed in the French *Petit Robert* dictionary (Robert, 1986).

Lexical corpus

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 which end in a consonant cluster + schwa (e.g. PORTE), since 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 non homophonic. The most frequent phonological structures were CVC (31%), CCVC (12%), and CV (11%). Table 1 displays the number of words as a function of number of letters and number of phonemes.

Table 1. Distribution of the 2,449 words as a function of number of letters and number of phonemes

Orthographic entries		Phonological entries	
number of letters	number of words	number of phonemes	number of words
1	0	1	26
2	57	2	354
3	252	3	1,023
4	623	4	580
5	863	5	357
6	515	6	101
7	123	7	8
8	16	8	0

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 posterior vowels [a] and [A]. This modification was motivated by the fact that the distinction is nearly completely lost in most current French dialects (Warnant, 1987; Léon, 1992). The phonological transcription is based on 15 vowels, 3 semi-vowels, 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. The phonetic symbols and their corresponding codes appear in Table 2a. Table 2b lists the orthographic codes used in the standard ASCII version for the letters with diacritics.

Table 2a. Characters coding the phonetic symbols in LEXOP

Vowels	Examples	LEXOP MAC/GENEVA codes	LEXOP ASCII codes	**Consonants**	Examples	LEXOP MAC/GENEVA codes	LEXOP ASCII codes
ɑ / a	bas, plat	a	4	p	père	p	p
e	blé	é	e	t	vite	t	t
ɛ	lait	è	E	k	sac	k	k
i	ville	i	i	b	robe	b	b
ɔ	mort	o	o	d	dos	d	d
o	peau	O	O	g	gare	g	g
u	route	u	u	f	fou	f	f
y	rue	y	y	s	sale	s	s
ø	deux	ø	2	ʃ	chat	∫	S
œ	peur	œ	9	v	rêve	v	v
ə	le	^	*	z	zèbre	z	z
ɛ̃	train	ê	5	ʒ	gel	j	Z
ɑ̃	vent	â	@	l	lent	l	l
ɔ̃	bon	ô	^	r	rose	r	r
œ̃	brun	û	1	m	main	m	m
Semi-vowels				n	nez	n	n
j	yeux	ï	j	ɲ	vigne	ñ	N
w	oui	ü	W	ŋ	swing	£	L
ɥ	huile	ÿ	w	x	loch	x	x

Note. The symbol “◊” (“#” in the ASCII version) is used to denote the null phoneme in the PHONPHO field.

Table 2b. Symbols coding the letters with diacritics

LEXOP MAC/GENEVA codes	LEXOP ASCII codes
ù	%
è	&
é	2
ë	5
ï	6
ü	8
à	@
â	A
ê	E
î	I
ô	O
û	U
ç	C

Segmentation principles

First, a segmentation algorithm was used to parse orthographic strings into graphemes by locating the letter string corresponding to each successive phoneme. To that end we employed a table of grapheme-phoneme correspondences compiled by Véronis (1986), to which a small number of entries were added manually.

A second operation consisted in parsing the phonological codes into onset (C_1), vowel (V) and coda (C_2). The C_1 and V units as well as the V and the C_2 units were then combined to compose the lead and the body-rime components, and the corresponding orthographic constituents were obtained by concatenating the relevant graphemic units.

The strict phonological segmentation principle adopted leads to some differences with an analysis of French body-rime correspondences recently reported by Ziegler, Jacobs, and Stone (1996). For example, in Ziegler et al.’s parsing, the letter U was considered as part of the vowel when following the letter G (as in the words GUIDE and GUISE), but was included in the onset when preceded by Q (as in the words QUITTE and QUOI). In contrast, in LEXOP, the letter U was always regarded as part of the onset when preceded by G or Q and followed by another vowel (see also Stanback, 1992, for a similar approach). A second difference with Ziegler et al. pertains to the treatment of semi-vowels. In their analysis semi-vowels were generally joined to the vowel, although it seems to depend on whether the semi-vowel was coded by orthographic vowels or consonants (e. g., the semi-vowel /j/ is considered part of the vowel for the word PIECE, but not for the word YACHT). Our parsing procedure followed standard phonological analyses of French in which semi-vowels are considered as consonants (e. g. Kaye & Lowenstamm, 1984)[1]. However, in a few cases (n= 79), the semi-vowel could not be distinguished orthographically from the vowel (e. g., OI and OIN in the words FROID and POINT are pronounced /wa/ and

[1] Note however that Kaye and Lowenstamm argue that semi-vowels and vowels may constitute complex nuclei, or diphthongs in a very restricted number of cases, of which the largest part corresponds to /wa/ and /we)/ when spelled OI or OIN.

/wE)/), and the semi-vowel was therefore considered as part of the vocalic unit. The same exceptional parsing also applied for two words with a pre-vocalic semi-vowel (POELE, MOELLE), and for four words with a post-vocalic semi-vowel (DRIVE, DRY, MILE, and PAYE). Mean lengths of C_1, V, and C_2 units resulting from the orthographic and phonological segmentations are given in Table 3.

Table 3. Mean length of C_1, V, and C_2 units

Unit	Number of letters	Number of phonemes
C_1	1.36	1.24
V	1.40	1.04
C_2	2.04	1.23

Computations

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.

The *frequency* (FRE) variables were computed as the number of occurrences of each ortho-phonological association in the word corpus. Correspondence frequencies are provided for C_1, V, C_2, C_1V and VC_2 units. In addition, for each word, the value of the least frequent grapheme-phoneme correspondence is recorded.

The *consistency* (CON) statistics were performed separately for orthographic to phonological mappings (OP) and phonological to orthographic mappings (PO). Ortho-phonological consistency was determined as the proportion of words containing a particular orthographic unit with the same pronunciation, relative to all words including the orthographic unit. In case of total consistency the ratio equates 1. For each lexical entry, type and token counts were carried out on C_1, V, C_2, C_1V and VC_2 units[2]. In addition, for each word, the value of the least consistent grapheme-phoneme and phoneme-grapheme correspondences is recorded.

Although in average, frequent correspondences are more consistent than rare correspondences, correspondence frequency and consistency are not necessarily related for a particular lexical entry. Several rare associations (as AON-/A)/ in the word PAON) are perfectly consistent, and several frequent associations (as ET-/E/ in the word GUET) are inconsistent (e.g. CET, NET).

Neighborhood size estimates were based on Coltheart et al.'s (1977) definition. *Orthographic* neighbors are words that can be generated from the base word by a single letter substitution. This definition was transposed to the *phonological* neighbors. Hence, phonological neighbors were words obtained by a single phoneme substitution. The *phonographic* neighbors designate the subset of all orthographic neighbors which are also phonological neighbors.

[2] C_1V and VC_2 counts were performed independently of the presence or absence of the C_1 and C_2 units in the words. As a consequence, for words starting with a vowel (V), the consistency values on CV correspond to the consistency of V computed by reference to the other CV units (and not by reference to the other V units as it is the case for the consistency values on V). The same logic applied to the VC_2 units for words ending with a vowel (without C_2).

The orthographic (ON), phonological (PN), and phonographic (PGN) neighborhood sizes were computed for each word in the corpus, using the same set of monosyllabic words as reference[3].

Finally, the phonographic neighborhood was analyzed as a function of the units shared with the base words. Neighbors can diverge from the base word on either C_1, V, or C_2 and thus share either VC_2 (*body/rime-neighbors*), C_1C_2 (*consonant neighbors*), or C_1V (*lead neighbors*), respectively.

Appendix A gives the full listing of the fields of the LEXOP database.

Descriptive statistics

Some analyses of the statistical properties of the LEXOP corpus have already been reported elsewhere (Peereman & Content, 1997), essentially for the neighborhood and consistency variables. In a first set of analyses (Peereman & Content, 1997), we showed the high consistency of OP correspondences on the subset of LEXOP consisting of CVC words, and we provided a detailed analysis of neighborhood space, demonstrating the numerical predominance of body-neighbors. In a second study (Content & Peereman, in preparation), we report a comparative analysis of these properties for English and French, including in addition sound-to-print statistics. It appeared that both French and English are relatively inconsistent on C_1, V and C_2 sound-to-print relations. Moreover, consistency did not increase significantly when larger size units (lead and rime) were involved in the analysis.

In the present Section, we describe statistical analyses of the LEXOP corpus that are relevant for stimulus selection. Obviously, two kinds of information are useful when searching for lexical items: the distributions of variables included in the experimental design, and the relations between these variables and other non-manipulated variables. Typically, researchers selecting materials along one specific dimension would need to control for potentially confounded factors of theoretical importance Hence, first we report descriptive statistics relative to the distributions of the LEXOP variables. In a second analysis, we examine the intercorrelations within the variable set.

Table 4 provides the mean and percentile values (10, 25, 50, 75, 90) for each of the variable considered in the LEXOP database. Not surprisingly, observations similar to those already reported in Peereman and Content (1997) can be made on the whole lexical corpus. More interesting for the user of the database is the information related to the distribution of the variables. As can be seen from Table 3, mean correspondence frequencies by type are very close to the medians, although generally slightly higher. A close look at the percentile values indicated quasi-symmetrical distribution from C_1, V, and C_2, as well as for the less frequent grapheme units. For CV and VC units, there are many words with very rare correspondences, and the distributions are positively skewed. More asymmetrical distributions are even observed for neighborhood variables, resulting from the fact that there are numerous words with very few neighbors. Regarding OP consistency, it clearly appears that most words include perfectly consistent C_1, V, C_2, C_1V, and VC_2 units. Even at the percentile 10 the values are already close to .80 or higher. These observations strengthen our previous finding of high consistency scores for French and show that the slight inconsistency is in fact due to a very limited set of words. Finally, except for C_1, PO consistency variables are distributed much more symmetrically than OP consistency variables.

Further analyses were carried out to examine the correlations between the LEXOP variables. These correlations should be taken into account when selecting stimuli to avoid confoundings with uncontrolled variables. As could be expected, a preliminary analysis indicated high correlations between type and token variables (from r=.19 to r=.97). To reduced the length of the correlation analyses between the LEXOP variables, we will thus exclusively consider type variables. The complete intercorrelation matrix, with p values higher than .01, is given in Table 5.

[3] The orthographic neighborhood size is slightly underestimated due to the exclusion of multisyllabic neighbors. Multisyllabic neighbors constitute only a very small part of the orthographic neighborhood of the monosyllabic words (6%). We report neighborhood statistics computed on LEXOP to keep the reference corpus identical for all analyses. An additional reason is that the facilitation effect of neighborhood size seems related to the number of phonographic neighbors, which include even fewer multisyllabic words (Peereman & Content, 1997).

Three main observations can be made. First, for each particular unit of analysis (C_1, V, C_2, CV, VC, grapheme-phoneme) there are high positive correlations between FREquency-CONsistency-and Neighborhood variables. For example, for C_1 and C_2 units, large correlations occur between OP consistency and PO consistency. There are also several strong correlations between variables on large size units (CV, VC) and variables bearing on one of the constituents (C_1 and V for CV, V and C_2 for V). For example, frequency of the correspondences on V units strongly correlates with frequency of the correspondences on CV units (r= .56). Second, as the neighborhood variables correspond to frequency estimations of particular orthographic and phonological patterns within subsets of the LEXOP corpus, almost all of the neighborhood variables correlate with the frequency variables. Third, although most correlations are positive, there are large negative correlations between the number of phonological neighbors and PO consistency on both C_2 and VC units (-.42 and -.49). This result follows from the fact that words with inconsistent C_2 units are often phonologically shorter than words with consistent C_2 units, and that short words have generally more numerous neighbors than long words.

Table 4. Mean and percentile values for each variable in the LEXOP database (by TYpe/by TOken)

Variables	Mean	P10	P25	P50	P75	P90
FREquency						
C_1	70/15,980	7/149	32/857	69/4,824	108/27,678	135/40,464
V	163/27,151	15/762	42/6420	167/19799	282/51,460	375/57,514
C_2	36/6,795	3/23	11/417	33/1,183	56/5,180	78/27,375
C_1V	6/1,238	1/4	2/22	4/192	10/741	16/2,650
VC_2	5/1,833	1/4	2/21	4/117	8/580	11/3,017
Least frequent (1)	86/1,037,587	11/24,155	25/117,063	60/554,805	129/1,541,240	209/2,431,773
OP CONsistency						
C_1	.96/.95	.90/.99	1/1	1/1	1/1	1/1
V	.90/.85	.76/.16	.86/.80	.99/1	1/1	1/1
C_2	.90/.92	.75/.90	.90/.98	1/1	1/1	1/1
C_1V	.94/.92	.80/.75	1/1	1/1	1/1	1/1
VC_2	.95/.95	.83/.99	1/1	1/1	1/1	1/1
Least frequent (1)	.43/.33	.11/.05	.26/.19	.44/.19	.59/.47	.74/.66
PO CONsistency						
C_1	.89/.88	.73/.41	.89/.99	1/1	1/1	1/1
V	.62/.66	.08/.03	.30/.35	.82/.94	.90/.97	.98/1
C_2	.58/.61	.11/.04	.26/.28	.61/.68	.92/1	1/1
C_1V	.67/.70	.18/.03	.40/.30	.78/.95	1/1	1/1
VC_2	.50/.52	.08/.01	.20/.07	.42/.53	.88/1	1/1
Least frequent (1)	.35/.28	.02/.01	.07/.03	.21/.25	.69/.33	.73/.83
Neighborhood						
ON	3.3/1,469	0/0	1/1	3/45	5/276	8/1379
PN	9.2/6,134	1/1	3/44	8/357	14/2612	20/17,285
PGN	2.3/1118	0/0	0/0	2/13	4/147	6/714
PGN C_1V	0.4/41	0/0	0/0	0/0	0/0	2/31
PGN VC_2	1.4/768	0/0	0/0	1/1	2/53	4/361
PGN C_1C_2	0.3/24	0/0	0/0	0/0	0/0	1/10

Note. (1) for grapheme-phoneme counts.

Unit Tables

In addition to the main database, the distribution includes two tables providing statistics on the different orthographic and phonological units of word segmentation (C_1, C_2, C_1V, V, VC_2). The OPTable file concerns orthography-to-phonology statistics, and the POTable file concerns phonology-to-orthography correspondences. Each includes data for C_1, C_2, C_1V, V, VC_2 units, successively.

The two sets of tables have the same structure: ORTHO lists the orthographic entries, PHON lists the phonological entries, TYPE and TOKEN provide the type and token frequencies of the correspondences. The fields ALLTYPE and ALLTOKEN give the type and token frequencies of alternative correspondences. Thus type consistency estimates could be computed as TYPE/(TYPE+ALLTYPE).

Note also that some of the OP and PO correspondences include null strings as some orthographic (phonological) units have no phonological (orthographic) counterparts.

References

Content, A., Mousty, P., & Radeau, M. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique, 90*, 551-566.

Content, A. & Peereman, R. (in preparation). *Quantitative analyses of orthography to phonology mapping in English and French*.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance* (Vol. VI, pp. 535-555). London: Academic Press.

Imbs, P. (1971). *Etudes statistiques sur le vocabulaire français. Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles*. Centre de Recherche pour un trésor de la langue française (CNRS), Nancy. Paris: Librairie Marcel Didier.

Kaye, J., & Lowenstamm, J. (1984). De la syllabicité. In F. Dell, D. Hirst, & J. R. Vergnaud (Eds.), *Forme sonore du langage. Structure des représentations en phonologie*, (pp. 123-159). Paris: Hermann.

Léon, P. (1992). *Phonétisme et prononciations du français*. Nathan.

Peereman, R. & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language, 37*, 382-410.

Peereman, R., & Content, A. (1999). LEXOP. A LEXical database providing Orthography-Phonology statistics for French monosyllabic words. To appear in Behavior Research Methods, Instruments, & Computers, 31, 376-379.

Robert, P. (1986). *Micro-Robert. Dictionnaire du Français primordial*. Paris: Dictionnaires Le Robert.

Robert, P. (1987). *Le Petit Robert. Dictionnaire alphabétique et analogique de la langue française*. Paris: Dictionnaires Le Robert.

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.

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

Warnant, L. (1987). *Dictionnaire de la prononciation française*. 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.

Appendix A. Fields of the LEXOP database

General information fields

ORTHO: ORTHOgraphic representation of the word
PHON: PHONological representation of the word
FREQ: word FREQuency *per million*, from Imbs (1977). Equals -1 when frequency not available
SCLASS: Syntactic class (NO: noun; AJ: adjective; VB: verb; AV: adverb; CO: conjunction; LO: locution; PN: pronoun; IN: interjection; AR: article; PR: preposition).
ORTHOCVC: ORTHOgraphic C_1-V-C_2 segmentation
ORTHOGRA: ORTHOgraphic segmentation in GRAphemes
PHONCVC: PHONological C_1-V-C_2 segmentation
PHONPHO: PHONological segmentation in PHOnemes

Frequency fields

FREC$_1$TY: correspondence FREquency on C_1, by TYpe
FREVTY: id. on V
FREC$_2$TY: id. on C_2
FRECVTY: id. on C_1V
FREVCTY: id. on VC_2
FREC$_1$TO: correspondence FREquency on C_1, by TOken
FREVTO: id. on V
FREC$_2$TO: id. on C_2
FRECVTO: id. on C_1V
FREVCTO: id. on VC_2
FRELTY: FREquency of the Least frequent grapheme-phoneme association, by TYpe
FRELTO: id., by TOken

Ortho-phonological Consistency fields

OPCONC$_1$TY: Ortho-Phonological CONsistency on C_1, by TYpe
OPCONVTY: id. on V
OPCONC$_2$TY: id. on C_2
OPCONCVTY: id. on C_1V
OPCONVCTY: id. on VC_2
OPCONC$_1$TO: Ortho-Phonological CONsistency on C_1, by TOken
OPCONVTO: id. on V
OPCONC$_2$TO: id. on C_2
OPCONCVTO: id. on C_1V
OPCONVCTO: id. on VC_2
OPCONLTY: Ortho-Phonological CONsistency of the Least consistent grapheme-to-phoneme correspondence, by TYpe
OPCONLTO: id., by TOken

Phono-orthographic Consistency fields

POCONC_1TY: Phono-Orthographic CONsistency on C_1, by TYpe
POCONVTY: id. on V
POCONC_2TY: id. on C_2
POCONCVTY: id. on C_1V
POCONVCTY: id. on VC_2
POCONC_1TO: Phono-Orthographic CONsistency on C_1, by TOken
POCONVTO: id. on V
POCONC_2TO: id. on C_2
POCONCVTO: id. on C_1V
POCONVCTO: id. on VC_2
POCONLTY: Phono-Orthographic CONsistency of the Least consistent phoneme-to-grapheme correspondence, by TYpe
POCONLTO: id., by TOken

Neighborhood size fields

ONTY: Orthographic Neighborhood size, by TYpe
ONTO: id., by TOken
PNTY: Phonological Neighborhood size, by TYpe
PNTO: id., by TOken
PGNTY: PhonoGraphic Neighborhood size, by TYpe
PGNTO: id., by TOken
PGNCVTY: PhonoGraphic Neighborhood size for C_1V (lead neighbors), by TYpe
PGNCVTO: id., by TOken
PGNVCTY: PhonoGraphic Neighborhood size for VC_2 (body-rime neighbors), by TYpe
PGNVCTO: id., by TOken
PGNCCTY: PhonoGraphic Neighborhood size for C_1C_2 (consonant neighbors), by TYpe
PGNCCTO: id., by TOken

Note: Numerical fields have a value of -9 when the information does not apply for a particular word. For example, the value of the C_1 consistency fields is equal to -9 when the words have no C_1 unit. The value -1 is used for numerical fields on token counts to signal that frequency data were not available (e. g., when there is no frequency estimates of the neighbors for a particular word).

Table 5. Intercorrelation matrix of the LEXOP variables (by type)

		FREQUENCY						OP CONSISTENCY						PO CONSISTENCY						NEIGHBORHOOD					
		C1	C2	V	CV	VC	LGP	C1	C2	V	CV	VC	LGP	C1	C2	V	CV	VC	LPG	ON	PN	PGN	CV	VC	CC
FRE	C1																								
	C2																								
	V		-20																						
	CV	**54**	-18	**56**																					
	VC	-14	**51**	25																					
	LGP	22		**44**	**41**	14																			
OP CON	C1	17	*7*			9	10																		
	C2		20			*8*																			
	V			29	19		17																		
	CV			15	18		13			**66**															
	VC		9			13			**46**	22	19														
	LGP	14	29		*9*	12	15	15		**32**	24														
PO CON	C1	**31**	*6*		9	*9*	18	**54**					15												
	C2	*7*				*-8*	18		**33**	*7*		19	-12												
	V	*6*	-12	**73**	**47**	19	**43**	*6*		17	*6*		13	*8*											
	CV		*-8*	**49**	**38**	15	**33**	10					11	29		**75**									
	VC	*8*	-20	*6*	10		15		20			17	-16		**71**	25	22								
	LPG	18	14	27	27	11	**69**	11		*8*	*6*		14	23	**37**	**51**	**45**	**41**							
N	ON	**44**	**32**	25	**47**	**43**	**35**					*-8*	22	12		25	16		28						
	PN	**44**	**31**	15	**41**	28	8			12	10	*-6*	29		**-42**		-11	**-49**	-15	**60**					
	PGN	**40**	**36**	24	**42**	**49**	**33**		11	12	12	*8*	22	14		23	15		24	**92**	**58**				
	CV	24	13	24	**41**	11	**35**		*6*	12	11		*7*		9	19	16		27	**59**	29	**62**			
	VC	30	**32**	19	**31**	**58**	20		*9*	*9*	10	10	19	10		21	4	*8*	16	**77**	**50**	**85**	22		
	CC	**31**	27		12	17	19	10	*8*				21	15				*8*	*9*	**49**	**37**	**54**	14	26	

Based on 1783 observations. Decimal point omitted. Correlation values with p>.01 have been removed. All p<.001 except correlation values in italics with .001<p<.01. Correlation values superior to .30 appear in bold