Behavior Research Methods, Instruments, & Computers
1996, 28 (3), 418-426

LEX I and II: Two databases of surface word forms for psycholinguistic research in Spanish

JULIO SANTIAGO, FERNANDO JUSTICIA, ALFONSO PALMA,
DOLORES HUERTAS, and NICOLÁS GUTIÉRREZ
University of Granada, Granada, Spain

Two databases of Spanish surface word forms are presented. Surface word forms are words considered as orthographically or phonologically specified without reference to their meaning or syntactic category. The databases are based on the productive written vocabulary of children between the ages of 6 and 10 years. Statistical and structural information is presented concerning surface word-form frequency, consonant–vowel (CV) structure, number of syllables, syllables, syllable CV structure, and subsyllabic units. LEX I was intended to aid in the study of reading processes. Entries were orthographic surface word forms; words were divided in their components following orthographic criteria. LEX II was designed for spoken language research. Accordingly, words were transcribed phonologically and phonological criteria were applied in extracting the internal units. Information about stress location was also provided. Together, LEX I and LEX II represent a useful tool for psycholinguists interested in the study of people acquiring Spanish as a first or foreign language and of Spanish-speaking populations in general.

Modern psycholinguistic research frequently requires statistical counts of linguistic items. The control or manipulation of word frequency is a classical example. It requires extracting a random sample of a language's lexicon and performing a frequency count. Two more recent examples of distributional variables that have an influence in language tasks are the number and frequency of lexical neighbors, defined as words that share most of their letters with a given written word (e.g., Grainger, 1990), and the frequency of printed syllables (Carreiras, Alvarez, & de Vega, 1993). The number of distributional variables to be controlled is increasing to the point that the most difficult part of the research is frequently the selection of a sufficient set of stimuli. This situation makes databases of linguistic materials a valuable tool, because they allow the user to select entries that simultaneously match the values of many different variables.

There is a lack of statistical tools for use in the study of Spanish-speaking populations. The most widely used word-frequency dictionary (Juilland & Chang-Rodriguez, 1964) is not recent and relies on written, mainly literary, material that may not reflect accurately the vocabulary that people use in everyday conversations. For research on sublexical units, there are frequency counts of printed syllables and bigrams (Alvarez, Carreiras, & de Vega, 1992a, 1992b), but no similar instruments have been published for tasks other than reading. Moreover, no tool of this type is available as a database. Therefore, the primary aim of the present work was to provide a linguistic database drawn on a sample of modern-day productive Spanish. The following paragraphs detail the scope of the information contained in it.

The database is primarily intended to aid in the study of sublexical units in the perception, production, and acquisition of Spanish. Psycholinguistic research on the importance of sublexical units is a rapidly growing field, both in speech production (e.g., Meyer, 1990, 1991) and reading (Prinzmetal, Treiman, & Rho, 1986; Treiman & Chafetz, 1987). Moreover, while there is debate about the relevance of some of these units in the English language (Cutler, Mehler, Norris, & Segui, 1986; Seidenberg, 1989), their importance for Spanish language processing is well established. In particular, the syllable seems to be an important unit in the perception of spoken and printed Spanish and other languages with clear syllabic structures (Carreiras et al., 1993; Cutler et al., 1986; Mehler, Dommergues, & Frauenfelder, 1981; Sebastian-Galles, Dupoux, Seguí, & Mehler, 1992). There is also an important set of studies that provide evidence for the psychological reality of units smaller than the syllable. The dominant view divides the syllable into two main components, the consonantal onset and the rime. The latter comprises the vowel and any following consonants. Within the rime, two more components are distinguished: the vowel nucleus and the coda, or final consonants. The evidence for these subsyllabic units is strong in the production of spoken English (see Treiman, 1989, for a review) and even stronger than that for alternative units in the perception of printed English (Treiman & Chafetz, 1987). As for syllable units, Spanish provides a stronger case than English for the reality of these subsyllabic units in different language tasks (Bradley, Sánchez-Casas, &

This work was supported in part by Grant AP91 42915819 from the Spanish Ministry of Science and Education to the first author. We gratefully thank Linda Siegel, Manuel Carreiras, Joe Lavallee, and two anonymous reviewers for their very helpful comments and suggestions on either form or content. Correspondence should be directed to J. Santiago, Dept. Psicologia Experimental y Fisiologia del Comportamiento, University of Granada, Granada 18001, Spain (e-mail: santiago@platon.ugr.es).

García-Albea, 1993; Pallier, Sebastián-Gallés, Felguera, Christophe, & Mehler, 1993; Sebastián-Gallés & Felguera, 1992).

The present databases contain information about a word's composition in terms of syllables, onsets, rimes, vowel nucleus, and codas, and provides certain more global indexes for the word, such as its frequency, its total number of syllables, and so on. Given our interest in sublexical structure information, we defined a word without any reference to its meaning or syntactic class. For example, the Spanish word CASA has three different meanings: "house," "to marry" in third person singular present indicative, and "to match" in the same conjugation. The three meanings and the two syntactic classes use a common word form, and as long as only their phonological and orthographic information is concerned, they are no longer different words. So, they are grouped together in the database. Words defined in this way will be called *surface word forms*, or just *surface forms* (SFs) from now on.

Spanish is a language with a shallow orthography, allowing accurate translation from printed words to sound following a fairly restricted set of rules. However, in the process of building the database, it was soon evident that a single database based on written SFs could not provide information relevant to both written and spoken Spanish. For example, the written SFs QUISO and KILO do not share any syllable. The first syllables of the two SFs differ from each other in terms of both the letters involved and the structure, the first one being a CVV syllable (where C stands for consonant and V for vowel) and the second being a CV syllable. However, when comparing the phonological forms of both words, /kíso/ and /kílo/, it turns out that the first syllables are identical in component phonemes and structure. Also, orthographic and phonological rules make some contrasting demands on the process of syllabification. For example, syllables in the written SF ADHERIR are divided as follows: AD, HE, and RIR. However, its phonological SF /aderíř/ is composed of the syllables /a/, /de/, and /riř/. Some more examples will be given in the section headed Syllabification Criteria.

These differences between orthographic and phonological structural information forced the creation of two twin databases that share most of their fields. LEX I was based on orthographic SFs, while LEX II was based on their corresponding phonological SFs. LEX II was generated from LEX I through a phonological transcription that assigned a single character to each individual phoneme. Surface forms in LEX I and LEX II were divided into syllables following slightly different criteria, as described above. The following sections describe common and contrasting aspects of the databases in greater detail.

The Sample

To increase the ecological validity of the information contained in the databases, we worked from the original sample of a recently published word-frequency count (Justicia, 1985). This sample reflected the vocabulary of modern-day productive Spanish among schoolchildren. Considering the fact that a large proportion of an individual's basic vocabulary is acquired during schooling, this database is probably also representative of current conversational vocabulary among adults.

The sample was obtained from compositions written by 2,166 6- to 10-year-old schoolchildren in southern Spain, who were asked to write about any topic of their choice. This procedure resulted in a sample of 255,711 words (tokens), comprising 12,281 different SF (types). The tokens were corrected for misspellings before being aggregated into types. Justicia's (1985) word count was based on the aggregation of types into basic forms, or lemmas. A lemma represents the whole set of words that derive from a single stem, for example, all conjugated forms of a verb. A total of 5,750 basic forms were obtained from the sample. In contrast, LEX I and LEX II are based on Justicia's (1985) original sample of tokens. What follows is a description of Justicia's sampling procedure.

The sampling used a stratified procedure. The sample was composed of several layers, defined by the factorial crossing of the following variables: age (6–10 years), origin (rural vs. urban), and province (Almería, Granada, Jaen, and Málaga).

The total number of sampled words was chosen on the basis of two main criteria: (1) the number of basic forms considered to be the normal productive vocabulary size for children between 6 and 10 years of age—between 5,000 and 6,000 basic forms, the upper limit given by Averril (1956). The obtained value was 5,750, as stated above; and (2) the average frequency of the basic forms, as indicated by the ratio between total number of tokens and basic forms. This ratio grows with the size of the sample, but follows a steadily accelerating function. That is, when the size of the sample is already large, a new sampled word will probably add to the total number of tokens but not to the total number of basic forms. When the ratio is high, the average repetition of basic forms is also high, and it can be said that the sample is an adequate reflection of the subjects' usual vocabulary. The requested value of this ratio was set to 45. The actual average repetition of basic forms in the sample was 44.47.

The number of subjects in the stratified sampling was set independently for each cell by the factors age, origin, and province. The criteria were threefold: (1) the total number of tokens needed, as specified by the criteria given in the last paragraph; (2) the variability in the verbal production of subjects in each cell (determined in a pilot study); and (3) the proportion of the subjects in each cell with respect to their source populations as defined by the statistics of the Spanish Ministry of Education. The final distribution of subjects per cell can be found in Justicia (1985).

Phonological Transcription

LEX II is based on a phonological transcription of the sample of orthographic SF, aimed to represent each phoneme by a single character. Spanish has a shallow orthography, which allowed us to rely on a small and clearly defined set of rules. Each of these rules is presented in Appendix A, which describes their target phonemes and conditions (letters that are acted upon), and gives some exam-

ples. Phonological symbols follow the conventions dictated by the Real Academia Española [Spanish Royal Academy of Language] (1991), as long as there is a standard ASCII character for it. Thus, in order to assure portability, the phonemes /ĉ/, /n̨/, and /ř/ were assigned the characters "c," "ñ," and "╞" in LEX II.

Syllabification Criteria

Syllabification criteria are clear in Spanish. Most speakers would agree on the syllable boundaries for a vast majority of words. This is so whether syllabification is performed by orthographic or phonological criteria. However, there are a few contrasting cases.

Basically, standard orthographic criteria are sensitive to letters that are not pronounced (H), and consider any union of a strong vowel (A, E, O) and a weak vowel (I, U) or of two weak vowels as a diphthong unless the weak vowel is marked with accentuation (á, é, í, ó, ú). So, for example, syllable boundaries for the word ANHIDRIDO (/anídrido/) are AN-HI-DRI-DO, and for the word JESUITA (/xesuíta/), JE-SUI-TA. Orthographic criteria also consider the letter X as a single unit, therefore rendering the boundaries EX-TA-SIS (/ékstasis/) and E-XI-TO (/éksito/).

In contrast, since phonology takes the phonological syllable as the single criterion, it does not take into account "silent" letters (H) and it does not consider some vowel clusters that are diphthongs by orthographic criteria as such. Phonologically, a diphthong is a sequence of two vowels where only one of them carries the syllabic stress and both are pronounced as a single unit. In some cases, vowel sequences that meet the orthographic criteria for diphthongs are not pronounced as a single unit. These cases are not considered diphthongs by phonological criteria. Applying these phonological criteria renders the following syllabic boundaries for the examples above: /a-ní-dri- do/, /de-sak-ti-bář/, and /xe-su-í-ta/. Finally, X is phonologically considered as a double phoneme /ks/. When these two phonemes are not followed by a vowel, they are clustered with the prior syllable, as in /eks-ta-sis/. When they are, the first phoneme is considered part of the last syllable, and the second joins the following one: /ék-si-to/.

Description of the Records

A record of LEX I will be presented first. Then the different fields in a record of LEX II will be described. Appendix B presents example records from LEX I and LEX II.

A record in LEX I contains the following information about a word: the orthographic SF, its CV structure, number of syllables, frequency, and a set of fields for each syllable. Words of up to seven syllables are included in the databases; this misses only a few composed multistem words.

The orthographic SF is written in upper case, and accent signs are missing. It introduces some ambiguity in the determination of stress location. There are Spanish words that are distinguished only by the position of the stress (contrastive accent). Spanish orthography codes this aspect of words unambiguously by a set of rules that control placement of the sign of accent on vowels, such that any pair of words that differ only in stress are distinguished by the presence of this sign in one of them. Thus, for example, the written SF AMO may correspond to "amo," with stress on the first vowel, or to "amó," in which the stress is on the second vowel. The problems that contrastive accent poses for the use of the present databases are discussed in the final section of this article.

CV structure is obtained by translating each vowel to the character V and each consonant to the character C. The frequency field contains the absolute number of times the SF appeared in the sample. Information about each syllable is contained in a set of fields. These sets are organized such that there is one set for each syllable position in the word, starting from the first up to the sixth, and then the last syllable. Thus the last set of fields contains information about the last syllable in the word, without regard to the word's length in syllables. For instance, a bisyllabic word would have entries in the first and the seventh sets of syllabic fields. Each set is composed of the following fields: the syllable string, a code reflecting its structure (described below), the onset and its length, the rime and its length, the vowel nucleus and its length, and the coda and its length.

Length refers to the number of letters composing the unit. For example, length of the onset in the syllable TRA would be 2. The structural code is a three-digit number reflecting the length of onset, vowel nucleus, and coda. For instance, the code for syllable TRA would be 210 and that for syllable PER, 111. The most frequent syllable structure (CV) will have the code 110.

A record in LEX II contains three new whole-word fields: the phonological SF, its CV structure, and a code identifying the syllable that carries the stress. This code takes a value of 1 if the stress is on the last syllable, 2 if it is on the next-to-last syllable, and so on. Stress-location coding was an a posteriori process, for it worked from the written SF contained in LEX I and not from the original sample of words. Because of the contrastive accent and the lack of accent signs, as described above, some phonological SFs could not be assigned a single stress-location code. In these cases, the stress code contains as many digits as there are possible stress locations in a word. For example, the code for the word AMO would be "12," the digits corresponding to the words "amó" and "amo," respectively. All other fields are the same in LEX II and LEX I, but their contents are based on the phonologically transcribed SF and its division into syllables following the phonological criteria.

Discussion

LEX I and LEX II provide a flexible database system for psycholinguistic research in Spanish populations. Stimulus material for many different kinds of experiments related to lower levels of analysis (orthographic and phonological) can be easily searched. Apart from the classic variable of frequency (defined here as SF frequency), words can be found that match a particular number of syllables, whose syllable components conform to particular structures, and so on. Both databases can also be useful to lin-

guists interested in statistical aspects of language. A frequency dictionary of printed Spanish syllables has already been generated from the material in LEX I (Justicia, Santiago, Palma, Huertas, & Gutiérrez, in press).

However, there are some aspects that may limit the validity of the information contained in the databases. The first concerns its generalizability to other dialects of Spanish. Dialectical variation affects the index of SF frequency. Some SFs are used more frequently in some dialects than in others. For example, in Spanish there are two words for "you": TU and USTED. In Spain (except for the Canary Islands), the latter is used in formal contexts only, and so its frequency is likely to be low. In contrast, it is broadly used in any conversational context by Latin American speakers of Spanish. So, its frequency in those dialects is likely to be high.

Clearly, differences in the frequency of some SFs are expected among different dialects. The frequency index can probably be confidently generalized to other Spanish dialects in Spain (for example, in northern Spain), but more differences are expected with Latin American Spanish. However, we believe that important changes in frequency will be apparent for only a relatively small subset of words, with most of them remaining within the same coarse frequency level. That is, if we use the index to divide the words in the sample in high-, medium-, and low-frequency entries, high correlations are to be expected among different dialects. This is an empirical problem that has to be addressed by comparing SF counts from different dialects.

It can also be argued that dialectical variations affect the sublexical information contained in the databases. For example, in southern Spain and Latin American Spanish /θ/ is pronounced as /s/ (e.g., the word CESTO is pronounced as /sésto/, instead of being pronounced as /θésto/). If these dialectical variations in pronunciation were included in the databases, its sublexical information could not be generalizable to other dialects. However, this is not the case. SFs in LEX II are the result of a phonological, not a phonetic, transcription. A phonological transcription does not take into account dialectical variations in the phonetic form of words. Phonemes are used instead of phones. Phonemes are units with contrastive value. That is, they discriminate among different semantic alternatives. At a phonological level, CASA ("house") and CAZA ("hunt") are differentiated by the contrast between /s/ and /θ/, independently of the fact that, in some dialects, both words may be realized in the same phonetic form /kása/. SFs in LEX I and II were regularized at two different levels: first, the written sample was corrected for misspellings before LEX I was generated; second, these orthographic SFs were phonologically transcribed in order to generate LEX II. By keeping sublexical information at a more abstract level than actual written or spoken performance, this process of regularization makes this information more generalizable to other Spanish dialects.

A second aspect that also affects the validity of the SF frequency index in both LEX I and LEX II is the contrastive accent. As stated above, the contrastive accent refers to the fact that in Spanish some words are distinguished from others just by the location of the stress. For example, CASO is a stress-ambiguous SF, because either the first or the second syllable can be stressed. Spanish orthography disambiguates these surface forms by means of the use of the accent symbol: "caso" versus "casó."

Contrastive accent is a problem for the frequency index in both databases because in LEX I accent symbols are missing (all words are written in upper case), which makes it impossible to separate the frequencies associated with "caso" and "casó." Because LEX II was generated from LEX I, it is affected by the same problem. We suggest, however, that the loss of contrastive accent is not an important problem for the orthographic database. In both reading and writing, knowledge about accent-sign placement rules can be analyzed as something that is independent of knowledge of the written form of words. Moreover, Spanish readers can perfectly identify and name aloud uppercase written words, despite the fact that they do not show the accent symbol. There have already been successful approaches to the study of the influence of subword units (such as syllables) in Spanish that did not take word stress into account (Carreiras et al., 1993). Carreiras et al. showed that syllable frequency had an effect in reading tasks, although stressed and unstressed syllables were pooled together in their estimations of syllable frequency.

The importance of this problem for the phonological database is more complex. It depends on whether phonological information, such as syllables, subsyllabic units, and segments, is closely linked to or independent from levels of stress. If metrical stress is represented independently of subword units, our databases can still be useful as directories of phonological information other than metrical stress. Current theories of language production disagree in this respect (compare, e.g., Dell, 1986, and MacKay, 1987, with Levelt, 1989).

Given the difficulty of discussing this issue on an a priori basis, it is better to take an empirical approach. Stress-ambiguous SFs (those that can be assigned more than a single stress pattern, with the alternatives producing legal Spanish words) in the sample were counted. Out of 12,281 types in our sample, 313 were stress ambiguous. This amounts to only 2.55%. That is, the vast majority of words in the sample were assigned a single stress pattern, in spite of the lack of orthographic accent symbols. This proportion is even smaller if we count the number of stress-ambiguous tokens (4,544) and make it relative to the total number of tokens in the sample (255,711); this amounts to only 1.78%. Moreover, both possible stress patterns were not present in the original sample with similar frequencies. Actually, in a large majority of cases, most of the tokens of a stress-ambiguous SF were consistent in their stress patterns. This argument is based on a consideration of the type of words that are stress ambiguous.

In most cases, the ambiguity is between two forms of the same verb (221 types out of 313, or 70.61%), although, when considering the frequencies of these types, it turns

out that stress-ambiguous words with alternatives of different syntactic classes are relatively more frequent (2,290 tokens out of 4,544, or 50.40%). The ambiguities between verb forms always fall into one of two categories, and in both cases it is easy to know the most frequent form in the sample. The first category includes ambiguities between the first-person singular present indicative (e.g., "amo," "I love") and the third-person singular past perfect indicative ("amó," "he loved"). Given the narrative nature of the children's written productions, past perfect was much more frequent than present. The second category includes ambiguities between the first-person singular future indicative (e.g., "trabajaré," "I will work") and the first- or third-person singular future subjunctive ("trabajare," "If I would work"). The future subjunctive is a highly infrequent and difficult verb conjugation that is used only in literary texts and very erudite conversational contexts. Even if our sample came from adult speakers, most of these tokens would be of the former type.

The above considerations show that the problem of contrastive accent is quite limited in the sample, and that even when the type is stress ambiguous, most (in many cases, all) of its frequency index comes from tokens with the same stress pattern, whatever it is. Because stress-ambiguous SFs are marked in LEX II by means of a multidigit stress code, the user can always identify the problematic items and what the alternative SFs that correspond to each of them are.

One final aspect that deserves comment is the usefulness of these tools for the study of adult language. It could be argued that the sample of words does not reflect the normal adult vocabulary. Although it is clearly true that an adult's vocabulary is different from that of a child, there are reasons to think that, at the level of usual vocabulary, they have more commonalities than differences. Adults acquire most of their basic vocabulary in school, and the word set that they use in everyday conversation is probably not significantly different from the one they used at the end of primary school, although the vocabulary of adults is undoubtedly more extended than children's. The cognitive world of adults is more complex, and this is perforce reflected in their vocabularies. Nevertheless, the nucleus of such a world (everyday actions, intentions, family and social relations, and so on) is very similar and shared among school children and adults. The more important differences are probably located in low-frequency words, where many new entries would be found in an adult lexicon. This means that a word sample from children may be representative of adults' most commonly used vocabulary. Unfortunately, no systematic comparisons between this sample and an adult sample of words can be offered. The standard Spanish word-frequency count is that by Juilland and Chang-Rodriguez (1964), but the sources of its sample (which uses mainly written texts), its date of publication, and especially the differences between their definition of word frequency and ours, make a direct comparison, at best, difficult.

Since LEX I and II are intended to help in the study of sublexical units in language performance, an important related question is whether or not the sublexical "vocabularies" of children and adults are similar. This question can be given an empirical answer. Alvarez et al.'s (1992a) syllable count is based on a sample of words taken from adult literature (newspapers, magazines, and books). Hence, it may be taken to represent the adult syllabary. Justicia et al. (in press) listed all syllables in LEX I, and found their frequency by adding syllable frequency and the frequency of the SFs that contained them. They then compared the resulting syllable-frequency dictionary with Alvarez et al.'s. The results of this comparison are summarized in Appendix C. Two aspects are especially noteworthy. First, syllables present in both dictionaries appeared in similar frequency ranks (Pearson's $r = .733$). Second, the number of syllables present in Justicia et al.'s dictionary but not in Alvarez et al.'s is more than twice that present in Alvarez et al.'s dictionary but not in Justicia et al.'s. This difference can be accounted for by the bigger sample size on which Justicia et al.'s dictionary was based. Although the difference in sample size causes us to take these results with caution, they suggest that the adult's syllabary is basically the same as that of the child. It makes good sense to think that the set of sublexical units is almost fixed during the first stages of training in reading, with the subsequent widening of vocabulary with age and education taking place mainly by virtue of the acquisition of novel combinations of those units.

To summarize, LEX I and LEX II can be considered as useful tools for the study of sublexical influences in language tasks both in children and adults. At the SF level, even if the adult vocabulary were quite different from and not comparable to that of a child, LEX I and LEX II would be valuable tools for the developmental psycholinguist and for those interested in the acquisition of Spanish as a second language.

Availability

LEX I and LEX II are implemented as DBASE IV databases and are available upon request. Please send $12 U.S. to cover diskette and mailing expenses.

REFERENCES

Alvarez, C. J., Carreiras, M., & De Vega, M. (1992a). Estudio estadístico de la ortografía castellana: (1) La frecuencia silábica [Statistical study of castilian orthography: (1) Syllable frequency]. *Cognitiva*, **4**, 75-106.

Alvarez, C. J., Carreiras, M., & De Vega, M. (1992b). Estudio estadístico de la ortografía castellana: (2) Frecuencia de bigramas [Statistical study of castilian orthography: (2) Bigram frequency]. *Cognitiva*, **4**, 107-125.

Averril, M. (1956). *La vida psiquica del escolar* [The schoolchild's psychology]. Buenos Aires: Paidós.

Bradley, D. C., Sánchez-Casas, R. M., & García-Albea, J. E. (1993). The status of the syllable in the perception of Spanish and English. *Language & Cognitive Processes*, **8**, 197-233.

Carreiras, M., Alvarez, C. J., & De Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory & Language*, **32**, 766-780.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory & Language*, **25**, 385-400.

DELL, G. (1986). A spreading activation theory of retrieval during sentence production. *Psychological Review*, **93**, 283-321.

GRAINGER, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory & Language*, **29**, 228-244.

JUILLAND, A., & CHANG-RODRÍGUEZ, E. (1964). *Frequency dictionary of Spanish words*. London: Mouton.

JUSTICIA, F. (1985). *El vocabulario usual del niño en el ciclo inicial y el ciclo medio de la EGB* [The usual vocabulary of children in the first five grades of primary school]. Granada: Universidad de Granada, Secretariado de Publicaciones.

JUSTICIA, F., SANTIAGO, J., PALMA, A., HUERTAS, L., & GUTIÉRREZ, N. (in press). La frecuencia silábica del español escrito por niños: Estudio estadístico [Syllable frequency: A statistical analysis of written productions by Spanish children]. *Cognitiva*.

LEVELT, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

MACKAY, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer-Verlag.

MEHLER, J., DOMMERGUES, J. Y., & FRAUENFELDER, U. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning & Verbal Behavior*, **20**, 298-305.

MEYER, A. (1990). The time course of phonological encoding in language production: The encoding of successive syllables in a word. *Journal of Memory & Language*, **29**, 524-545.

MEYER, A. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory & Language*, **30**, 69-89.

PALLIER, C., SEBASTIÁN-GALLÉS, N., FELGUERA, T., CHRISTOPHE, A., & MEHLER, J. (1993). Attentional allocation within the syllabic structure of spoken words. *Journal of Memory & Language*, **32**, 373-389.

PRINZMETAL, W., TREIMAN, R., & RHO, S. (1986). How to see a reading unit. *Journal of Memory & Language*, **25**, 461-475.

REAL ACADEMIA ESPAÑOLA (1991). *Esbozo de una nueva gramática de la lengua española* [Toward a new Spanish grammar]. Madrid: Espasa-Calpe.

SEBASTIÁN-GALLÉS, N., DUPOUX, E., SEGUÍ, J., & MEHLER, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory & Language*, **31**, 18-32.

SEBASTIÁN-GALLÉS, N., & FELGUERA, T. (1992). Detección de fonemas en ataques y codas silábicos [Detection of phonemes in syllabic onsets and codas]. *Cognitiva*, **4**, 173-191.

SEIDENBERG, M. S. (1989). Reading complex words. In G. N. Carlson & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 53-105). Dordrecht: Klüwer.

TREIMAN, R. (1989). The internal structure of the syllable. In G. N. Carlson & M. K. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 27-52). Dordrecht: Klüwer.

TREIMAN, R., & CHAFETZ, J. (1987). Are there onset- and rime-like units in printed words? In M. Coltheart (Ed.), *The psychology of reading: Attention and performance XII* (pp. 281-298). Hillsdale, NJ: Erlbaum.

(*Continued on next page*)

APPENDIX A

Phonological Transcription Rules Used to Produce the Words in LEX II

Notational Conventions

1. The target letter or letter cluster is given in bold type.
2. An underline following a letter means "followed by"; one preceding it means "preceded by." For instance, "**c**_e" stands for "c when followed by e."
3. A group of alternative letters is enclosed by squared brackets. For example, "**m**_[p,b,f,m]" means "m when followed by p, b, f, or m."
4. The character "#" denotes a syllable boundary, and "*" denotes a word boundary.
5. Whenever they are conflictive, the rules are given in order of importance, both within and between rows. For example, "**gu**_[e,i]" is given before "**g**_[a,o,u]," indicating that it should be applied first. Also, rules for /n/ are given before those for /m/, because sometimes the character m corresponds to the phoneme /n/, whereas the opposite is not true.

Special Letters

The letter h, when not in the cluster ch, is never pronounced in Spanish. It is therefore not transcribed. The letter x corresponds to two different phonemes: /ks/.

Note

Place of accentuation, as indicated in the phonological transcriptions in the table, is not included in the phonological information in LEX II.

Phoneme	Conditions	Examples	Transcription
/a/	**a**, **á**	adiós	/adiós/
/b/	**b**, **v**	barca, avalar	/bářka/, /abalář/
/θ/	**c**_[e,i], **z**	cesto, zapato	/θésto/, /θapáto/
/ĉ/[1]	**ch**	chiste	/ĉíste/
/d/	**d**	dar	/dář/
/e/	**e**, **é**	ser	/seř/
/f/	**f**	fuerte	/fuéřte/
/g/	**gu**_[e,i], **g**_[a,o,u,ü]	gato, guerra	/gáto/, /géřa/
/i/	**i**, **í**, **y**_[#,*], **y**_[not vowel]	hilo, hoy	/ílo/, /ói/
/x/	**g**_[e,i], **j**	gesto, juez	/xésto/, /xuéz/
/k/	**k**, **c**_[a,o,u], **c**_[#,*]	kilo, cal, caćtus	/kílo/, /kál/, /káktus/
/l/	**l**	loro	/lóro/
/n/	**n**, **m**_[p, b, f, m]	noche, campo	/nóĉe/, /kánpo/
/m/	**m**	mañana	/maŋána/
/ŋ/[1]	**ñ**	niño	/níŋo/
/o/	**o**, **ó**	oro	/óro/
/p/	**p**	pelo	/pélo/
/ř/[1]	**rr**, *_**r**, [liquid, nasal]_**r**_[vowel]	carro, rato, alrededor	/kářo/, /řáto/, /alřededóř/
/r/	[vowel]_**r**_[vowel], #_[p,t,c,k,b,d,g,f]_**r**	cara, compra, fresa, piedra, cráter	/kára/, /kónpra/, /frésa/, /piédra/, /kráteř/
/s/	**s**	sal	/sál/
/t/	**t**	atar	/atař/
/u/	**u**, **ú**, **ü**	usar, útil, cigüeña	/usář/, /útil/, /θiguéŋa/
/y/	**ll**, **y**_[vowel]	llave, yo	/yábe/, /yó/

[1]These phonemes have been assigned the following ASCII characters in the database: /ĉ/→c; /ŋ/→ñ; /ř/→⌐.

APPENDIX B

This Appendix comprises two records for the word CHIQUILLO, the first from LEX I and the second from LEX II. The whole set of subsyllabic fields is shown only for the first syllable, but is included in the databases for all other syllables of the word.

LEX I

FIELDS			CONTENT EXAMPLE
Written word			CHIQUILLO
Orthographic CV structure			CCVCVVCCV
Number of syllables			3
Absolute frequency			3
Syllabic fields	First Syllable	Syllable	CHI
		Structure code	210
		Onset	CH
		Length	2
		Rime	I
		Length	1
		Vowel Nucleus	I
		Length	1
		Coda	.
		Length	0
	Second Syl.	Syllable fields	QUI
	Third Syl.	Syllable fields	.
	Fourth Syl.	Syllable fields	.
	Fifth Syl.	Syllable fields	.
	Sixth Syl.	Syllable fields	.
	Last Syl.	Syllable fields	LLO

(*Continued on next page*)

APPENDIX B (Continued)

LEX II

FIELDS			CONTENT EXAMPLE
Written word			CHIQUILLO
Orthographic CV structure			CCVCVVCCV
Phonologically transcribed word			/ĉikiyo/
Phonological CV structure			CVCVCV
Number of syllables			3
Stress location			2
Absolute frequency			3
Syllabic fields	First Syllable	Syllable	/ĉi/
		Structure code	110
		Onset	/ĉ/
		Length	1
		Rime	/i/
		Length	1
		Vowel Nucleus	/i/
		Length	1
		Coda	.
		Length	0
	Second Syl.	Syllable fields	/ki/
	Third Syl.	Syllable fields	.
	Fourth Syl.	Syllable fields	.
	Fifth Syl.	Syllable fields	.
	Sixth Syl.	Syllable fields	.
	Last Syl.	Syllable fields	/yo/

APPENDIX C

Comparison Between Alvarez et al.'s (1992a) and Justicia et al.'s (in press) Syllable-Frequency Dictionaries

	Alvarez et al. (1992a)	Justicia et al. (in press)
Sample size (tokens)	24,967	255,711
Total number of syllables	959	1,148
Number of syllables not present in the other dictionary	127	316
Number of syllables in common	832	
Frequency rank correlation	0.733	

(Manuscript received June 23, 1994;
revision accepted for publication February 3, 1995.)