second
language
research

Second Language Research
28(1) 103–112
© The Author(s) 2012
Reprints and permission: sagepub.
co.uk/journalsPermissions.nav
DOI: 10.1177/0267658311432199
slr.sagepub.com
SAGE

Short research note: The Beginning Spanish Lexicon: A web-based interface to calculate phonological similarity among Spanish words in adults learning Spanish as a foreign language

Michael S. Vitevitch, Melissa K. Stamer, Douglas Kieweg
University of Kansas, USA

Abstract

A number of resources provide psycholinguistic researchers with information about the words that the typical child or adult knows in a variety of languages. What is currently not available is a resource that provides information about the words that a typical adult learning a foreign language knows. We created such a resource for Spanish: The Beginning Spanish Lexicon. The present report describes the words contained in this web-accessible resource, and the information about those words provided by the interface. This information is freely accessible at: http://www.people.ku.edu/~mvitevit/BegSpanLex.html

Keywords

on-line database, Spanish, neighborhood density, foreign language learner

I Introduction

A number of existing resources provide psycholinguistic researchers with information about the words that the typical adult knows in a variety of languages, including English (Kučera and Francis, 1967; Vitevitch and Luce, 2004), Spanish (BuscaPalabras, see Davis and Perea, 2005), Dutch, and German (CELEX, see Baayen et al., 1995), among

Corresponding author:
Michael S. Vitevitch, Department of Psychology, University of Kansas, 1415 Jayhawk Boulevard, Lawrence, KS 66045, USA
Email: mvitevit@ku.edu

other languages. Similar resources also provide information about the words that children typically know (Storkel and Hoover, 2010). In the present report, we describe the Beginning Spanish Lexicon; a resource that is equally important to – yet conspicuously absent from – the research landscape, namely, information about the words that a typical adult learning a foreign language knows.[1] This absence is especially surprising given the growth of research in this and related areas (e.g. Kroll and De Groot, 2005).

II Words in the Beginning Spanish Lexicon

The 3,854 words contained in the Beginning Spanish Lexicon were obtained from the glossary of the first-year Spanish textbook, *¿Sabías que...?: Beginning Spanish* (VanPatten et al., 2004). The use of dictionaries and similar sources is a well-established – though not perfect – method of estimating the contents of the vocabulary of the average speaker (Luce and Pisoni, 1998). No proper nouns (e.g. country names, language names) were included in the Beginning Spanish Lexicon. With regard to verbs, only the infinitival form of each verb appeared in the glossary, and so only the infinitival form appears in the lexicon. That is, verb conjugations were not included. For example, the verb *decir* appears in the Beginning Spanish Lexicon, but its conjugates *digo, dices, dice*, and so on do not.

The words in the Beginning Spanish Lexicon were transcribed with a computer-readable phonemic font (transcriptions performed by the second author). The computer-readable phonemic font uses the following ASCII characters: p, t, k, b, d, g, C, s, x, f, T, n, m, 3, l, y, r, 5, 6, I, e, a, o, u, w, and j to represent the phonemic inventory in Spanish. A table containing the computer-readable characters and the corresponding Spanish phonemes – represented with the symbols of the International Phonetic Alphabet (IPA) – can be found in Appendix 1 (and on the website for the Beginning Spanish Lexicon). Researchers who have used a similar ASCII-transcription with English words to compute neighborhood density (e.g. Storkel and Hoover, 2010) or phonotactic probabilities (e.g. Vitevitch and Luce, 2004) should note that the ASCII characters x and r are used to represent different phonemes in English and in Spanish. These differences have been highlighted in Appendix 1.

To facilitate comparison with existing psycholinguistic resources (e.g. *LEXESP*; Sebastián Gallés et al., 2000) the transcriptions used in the Beginning Spanish Lexicon are Castilian Spanish. Therefore, a /θ/ was used to reflect particular sounds as found in Castilian Spanish. For example, in Castilian Spanish a /θ/ would be used in *cazar* and a /s/ would be used in *casar*. The /θ/ in Castilian Spanish is used for the following letters/letter strings: *z, ce,* and *ci*, and is represented with a T in the computer-readable transcription. Similarly, the letter *v* is represented with the phoneme /b/ and b in the computer-readable transcription. Finally, given the regularity of lexical stress in Spanish (stress typically occurs on the penultimate syllable of a word) stress markers were not included in the transcription.

We recognize that the Beginning Spanish Lexicon has limitations,[2] and is an imperfect approximation of the vocabulary a typical learner will acquire after one year of studying Spanish in an academic setting at the college level. Despite these shortcomings we believe that the Beginning Spanish Lexicon better captures such knowledge than simply using the most frequent words from a native-language resource (see Davies and

Face, 2006). Furthermore, as described below, the Beginning Spanish Lexicon also provides information that is not available in other presently available resources (e.g. Davies, 2006).

III Information provided by the Beginning Spanish Lexicon

Researchers provide phonemic representations as input to the Beginning Spanish Lexicon using the computer-readable phonemic font described above. For example, the word *cazar* would be phonologically represented and input to the web-interface as kaTa5. Given the relatively transparent mapping between phonology and orthography in Spanish, this convention should not impose too much of an inconvenience to users. Instead, we believe this convention allows more flexibility for researchers to investigate a wider variety of research questions, especially questions regarding the phonological influence of one language on another (e.g. Vitevitch, 2011), or questions regarding cognates. Using phonemic representations also enables researchers to easily obtain information about specially constructed nonsense words that could be used to investigate knowledge of the phonotactic rules of Spanish (for a study that assessed phonotactic knowledge in English, see Vitevitch et al., 1997).

Another example of how phonemic inputs provide flexibility for researchers can be found in a study examining the acquisition of real Spanish words that might be just beyond what the typical language learner knows: Are novel words that resemble many known Spanish words easier or more difficult to acquire than novel words that resemble few known Spanish words? By using phonemic inputs to the Beginning Spanish Lexicon one could examine this question with specially constructed nonwords as has been done in previous experiments performed with English speakers (e.g. Storkel et al., 2006) or with real Spanish words (e.g. Stamer and Vitevitch, 2011). Basing such estimates on native language resources might over-estimate the lexical knowledge of the typical language learner, and would not identify *which* words in the native lexicon a typical language learner knows.

The Beginning Spanish Lexicon provides as output:

- the phonological representation (in the computer-readable transcription) that was initially entered;
- the orthographic representation (i.e. how it is spelled) of the word if it is contained in the Lexicon using the Roman alphabet;
- the number of phonemes in the word;
- the number of syllables in the word (n.b. there are no syllable markers in the phonological representations) if it is contained in the Lexicon;[3]
- the number of words that are phonologically similar (based on the criteria described below) to the input string; and
- the words from the Lexicon that are phonologically similar to the input string.

Two words are said to be phonologically similar (or are ‘phonological neighbors’) if the addition, deletion or substitution of any phoneme in one word forms the other word. This metric is commonly used in psycholinguistic research to assess phonological similarity (Luce and Pisoni, 1998). For example, substituting a phoneme in the initial position of the Spanish word *pato* forms the neighbors *dato, gato*, and *rato*. Substituting a phoneme in the third position of the word *pato* forms the neighbors *palo, paso*, and *pavo.* Substituting a phoneme in the fourth position of the word *pato* forms the neighbor *pata.* Finally, adding a phoneme to the word *pato* forms the neighbor *plato*. Although other Spanish words might exist that are neighbors of *pato,* none of those words were found in the Beginning Spanish Lexicon (i.e. someone just beginning to learn Spanish does not know those words).

The number of words that are phonologically similar to a target word is referred to in the psycholinguistic literature as 'neighborhood density' (Luce and Pisoni, 1998). Words with many similar sounding words are said to have a 'dense neighborhood', whereas words with few similar sounding words are said to have a 'sparse neighborhood'. (A median split is often used to categorize words as either dense or sparse, but other conventions can be used; e.g. Storkel, 2004). Numerous studies have shown that neighborhood density influences a variety of language-related processes including learning novel (English) words by children and adults (Storkel, 2002; Storkel et al., 2006), learning novel words in a second language (Stamer and Vitevitch, 2011), recognizing spoken words (for English, see Luce and Pisoni, 1998; for Spanish, see Vitevitch and Rodríguez, 2005), producing spoken words (for English, see Vitevitch, 1997; for Spanish, see Vitevitch and Stamer, 2006, 2009), and storing/retrieving words from short-term memory (Roodenrys et al., 2002).

Because neighborhood density in the Beginning Spanish Lexicon is computed from the set of words that a typical learner knows after one year of Spanish in an academic setting at the college level, the neighborhood density value may differ from the neighborhood density value for the same word obtained from a resource based on native-speakers of Spanish (e.g. BuscaPalabras, see Davis and Perea, 2005). Analyses of children and adults acquiring English as their native language (Charles-Luce and Luce, 1990) shows that both dense and sparse words gradually acquire more neighbors over time (but a dense word for a child does not become sparse in the adult lexicon, nor vice versa). Given the gradual increase in the size of phonological neighborhoods over time one might attempt to mathematically extrapolate from the lexicon of a native speaker of Spanish to the lexicon of someone beginning to learn Spanish to determine how many words are phonologically similar to a given word. Such an approach, however, will not provide information about which words in the native-Spanish lexicon the language learner knows (which might be useful for experiments that involve priming, etc.).

IV How to use the web-based interface

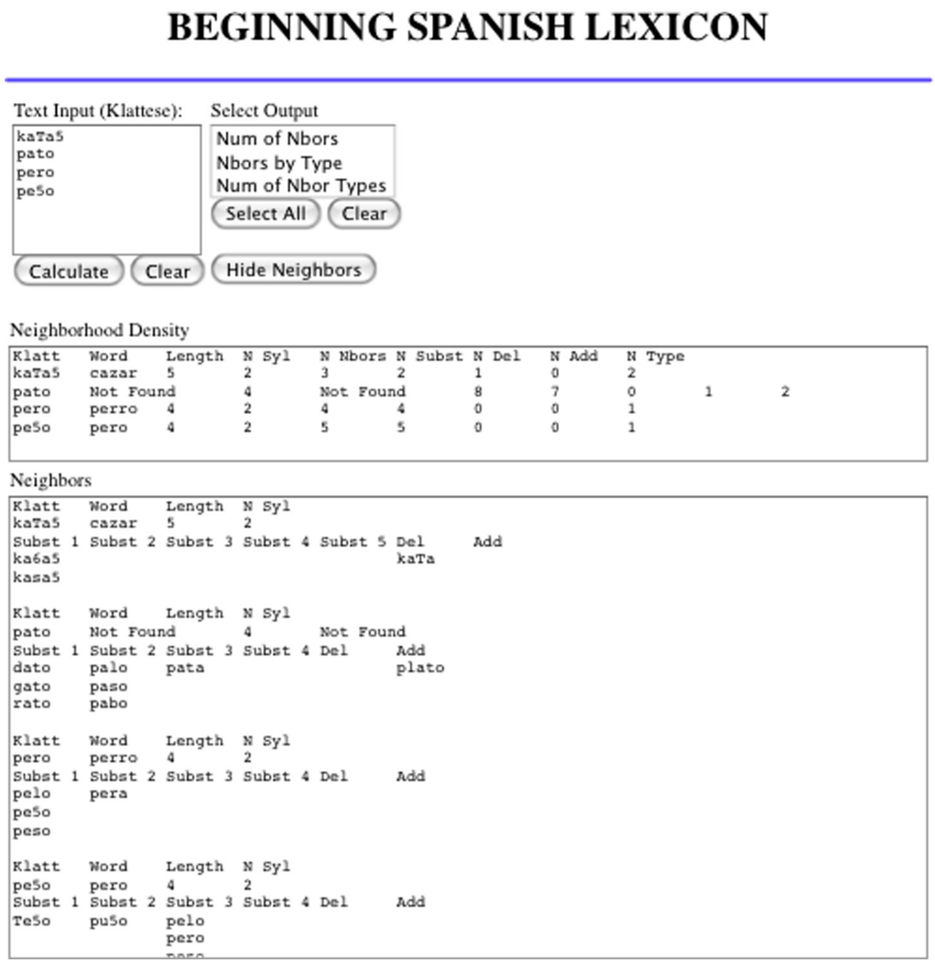
The Beginning Spanish Lexicon can be found at the following URL: http://www.people.ku.edu/~mvitevit/BegSpanLex.html. To calculate the neighborhood density of a word, the user should enter a single string of letters following the conventions of the computer-readable transcription (see Appendix 1) into the **Text Input (Klattese)** box. If the user wishes to enter more than one word, then each word should be entered on its own line followed by a hard return or a space (n.b. commas should not be used to separate words). Multiple words can be copied and pasted from word-processing programs if the words are separated by hard returns or spaces. When entering a word to be analysed, it is important to remember that the letter string must be the computer-readable – and case-sensitive – phonological transcription (i.e. a ‘C’ represents the ‘ch’ sound but a ‘c’ does not exist in the computer readable transcription of Spanish phonemes). Recall that orthographic representations should never be entered.

In the **Select Output** box, the user can select from several options including the number of neighbors, **Num of Nbors** (representing neighborhood density), and the types of neighbors that are found, **Nbors by Type**. ‘Type of neighbor’ refers to whether a single phoneme was substituted (**N Subst**), added (**N Add**), or deleted (**N Del**) to form a neighbor. Multiple fields can be selected by holding down the Ctrl or Shift Key while making a selection, or by clicking the ‘Select All’ button to choose all fields. Clicking the ‘Show Neighbors’ button will allow the user to see which words from the lexicon are phonologically related to the entered string. (Note that when the ‘Show Neighbors’ button is clicked, the button then shows the option to ‘Hide Neighbors’ as shown in Figure 1.) After clicking the ‘Calculate’ button the selected information (number of neighbors, type of neighbors, number of neighbor types, and the neighbors if desired) will appear in the window labeled ‘Neighborhood Density’ and, if ‘Show Neighbors’ was selected, in the window labeled ‘Neighbors’.

Figure 1 shows an example of the Beginning Spanish Lexicon using the words *perro* [pero] (English ‘dog’) and *pero* [pe5o] (English ‘but’) as input. Recall that the orthographic form of the words (*perro* and *pero*) is not entered as input. Instead, the computer readable phonemic transcription (the character strings in the square brackets in the examples above) is used as input. The ‘Select All’ button was clicked to provide as output information about neighborhood density (**Num of Nbors**) and the type of neighbors (**Nbors by Type** and **Num of Nbor types**). The ‘Show Neighbors’ button was also clicked (now showing the option to ‘Hide Neighbors’), resulting in the appearance of the ‘Neighbors’ window (which contains the neighbors sorted by the type of neighbor and the location in the word that a neighbor was formed).

In the ‘Neighborhood Density’ window, the computer readable transcription is shown (labeled as ‘Klatt’), as well as the orthography of the word (labeled as ‘Word’). Note that the orthography of the word will only appear if the word is among the 3,854 words in the Beginning Spanish Lexicon. If the word is a nonsense word, or a Spanish word that is not likely to be known by the typical adult with approximately one year of experience with Spanish (e.g. *pato*), then the phrase ‘Not Found’ will appear in the ‘Word’ column (however, neighborhood density information will still be calculated, enabling a researcher to use nonwords, advanced Spanish words, English words, etc. in an experiment). The number of phonemes (in the column labeled **Length**) and the number of syllables (in the column labeled **N Syll**) in the word is then displayed (n.b. the number of syllables will only be displayed if the word is found in the lexicon). The total number of phonological neighbors (i.e. neighborhood density) is then displayed in the column labeled **N Nbors**; the word *pero* has 4 neighbors, and the word *perro* has 5 neighbors. The number of neighbors that are formed by the substitution, deletion, and addition of a phoneme is then displayed, with **N Type** indicating the number of different types of neighbors formed. In the case of *pero* and *perro*, only the substitution of a phoneme forms phonological neighbors, so **N Subst** is the same number as **N Nbors**, and **N Type** is 1.

In the ‘Neighbors’ window, the same information provided in the ‘Neighborhood Density’ window is displayed again (although in a slightly different format). In



BEGINNING SPANISH LEXICON

Text Input (Klattese):

```
kaTa5
pato
pero
pe5o
```

Neighborhood Density

```
Klatt   Word    Length  N Syl   N Nbors N Subst N Del   N Add   N Type
kaTa5   cazar   5       2       3       2       1       0       2
pato    Not Found       4       Not Found       8       7       0       1       2
pero    perro   4       2       4       4       0       0       1
pe5o    pero    4       2       5       5       0       0       1
```

Neighbors

```
Klatt   Word    Length  N Syl
kaTa5   cazar   5       2
Subst 1 Subst 2 Subst 3 Subst 4 Subst 5 Del     Add
ka6a5                                   kaTa
kasa5

Klatt   Word    Length  N Syl
pato    Not Found       4       Not Found
Subst 1 Subst 2 Subst 3 Subst 4 Del     Add
dato    palo    pata                    plato
gato    paso
rato    pabo

Klatt   Word    Length  N Syl
pero    perro   4       2
Subst 1 Subst 2 Subst 3 Subst 4 Del     Add
pelo    pera
pe5o
peso

Klatt   Word    Length  N Syl
pe5o    pero    4       2
Subst 1 Subst 2 Subst 3 Subst 4 Del     Add
Te5o    pu5o    pelo
                pero
```

Figure 1. An image of the Beginning Spanish Lexicon with examples described in the text

addition, the actual words (in the computer readable phonemic transcription) from the Beginning Spanish Lexicon that are neighbors of the input string are displayed. Note that the words are categorized according to the type of neighbor the word is (i.e. a substitution in the various phoneme positions of the word, deletions of a phoneme, or the addition of a phoneme). For the word *perro* [pero] ‘dog’, all of the neighbors are formed by the substitution of a phoneme in the third position (*pelo* [pelo] ‘hair’, *pero* [pe5o] ‘but’, *peso* [peso] ‘weight’, or the currency used in Mexico) or fourth position (*pera* [pera] ‘pear’) of the word. Comparable information for the word *pero* [pe5o] can be viewed by using the scroll bar on the right side of the ‘Neighbors’ window. Alternatively, all of the information in the ‘Neighborhood Density’ and ‘Neighbors’ windows can be copied and pasted to a word processing program, spreadsheet, or other application to facilitate viewing, selection, manipulation, etc. of the output information.

We believe that the resource described here and made available to the scientific community has great potential to stimulate new research questions, and to facilitate the transition from basic research to practical and clinical applications. The information provided by the Beginning Spanish Lexicon might prove useful not only to researchers examining the influence of phonological neighbors on the recognition and production of spoken words (e.g. Vitevitch and Rodríguez, 2005; Vitevitch and Stamer, 2006, 2009), but may also prove useful to Speech-Language Pathologists who might wish to contrast or emphasize specific sounds or words in a therapy session with a bilingual client. Emphasizing the phonological differences among phonological neighbors might also have pedagogical value to second-language instructors (e.g. Stamer and Vitevitch, 2011), making the Beginning Spanish Lexicon a useful tool in the classroom as well.

Acknowledgements

This research was supported in part by grants from the National Institutes of Health to the University of Kansas through the Schiefelbusch Institute for Life Span Studies (National Institute on Deafness and Other Communication Disorders (NIDCD) R01 DC 006472), the Kansas Intellectual and Developmental Disabilities Research Center (National Institute of Child Health and Human Development P30 HD002528), and the Center for Biobehavioral Neurosciences in Communication Disorders (NIDCD P30 DC005803).

Notes

1. It is unlikely that the Beginning Spanish Lexicon includes every word that a student knows, and it is unlikely that all language students know every word found in the Beginning Spanish Lexicon (i.e. no student learns everything that is taught). To prevent redundancy in the text we will simply use the word 'know,' although the reader should understand that we do not imply anything so definitive in terms of a learner's vocabulary and mean something like 'is likely to know' or 'probably knows.'
2. The Beginning Spanish Lexicon does not provide the frequency with which words occur in the language (a.k.a. *word frequency*). However, several word-frequency counts presently exist for Spanish (e.g. Davies, 2006; Davis and Perea, 2005).
3. To obtain a measure of word-length in terms of the number of letters in the word the user can easily 'cut' the text-based output provided by the Beginning Spanish Lexicon and 'paste' it into an Excel spreadsheet. The character-length function in Excel (i.e. =LEN(<desired cell>)) can then be used to automatically compute the number of letters in the word.

References

Baayen RH, Piepenbrock R, and Gulikers L (1995) *CELEX-2* [CD-ROM]. Philadelphia, PA: University of Pennsylvania.

Charles-Luce J and Luce PA (1990) Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language* 17: 205–15.

Davies M (2006) *A frequency dictionary of Spanish: Core vocabulary for learners.* New York: Routledge.

Davies M and Face TL (2006) Vocabulary coverage in Spanish textbooks: How representative is it? In: Sagarra N and Toribio AJ (eds) *Selected proceedings of the 9th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla Proceedings Project, 132–43.

Davis CJ and Perea M (2005) BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods* 37: 665–71.

Kroll JF and De Groot AMB (eds) (2005) *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press.

Kučera H and Francis WN (1967) *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Luce PA and Pisoni DB (1998) Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19: 1–36.

Roodenrys S, Hulme C, Lethbridge A, Hinton M, and Nimmo LM (2002) Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28: 1019–34.

Sebastián Gallés N, Martí Antonín MA, Carreiras Valiña MF, and Cuetos Vega F (2000) *Lexesp: Léxico informatizado del español [*Computerized Spanish vocabulary*]* [CD-ROM]. Barcelona: Edicions de la Universitat de Barcelona.

Stamer MK and Vitevitch MS (2011) Phonological similarity influences word learning in adults learning Spanish as a foreign language. *Bilingualism: Language and Cognition*. DOI: 10.1017/S1366728911000216.

Storkel HL (2002) Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language* 29: 251–74.

Storkel HL (2004) Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research* 47: 1454–68.

Storkel HL and Hoover JR (2010) An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods* 42: 497–506.

Storkel HL, Armbruster J, and Hogan TP (2006) Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research* 49: 1175–92.

VanPatten B, Lee JF, and Ballman TL (2004) *¿Sabías que...?: Beginning Spanish* . 4th edition. New York: McGraw Hill.

Vitevitch MS (1997) The neighborhood characteristics of malapropisms. *Language and Speech* 40: 211–28.

Vitevitch MS (2011) What do foreign neighbors say about the mental lexicon? *Bilingualism: Language and Cognition*. DOI: 10.1017/S1366728911000149.

Vitevitch MS and Luce PA (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers* 36: 481–87.

Vitevitch MS and Rodríguez E (2005) Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders* 3: 64–73.

Vitevitch MS and Stamer MK (2006) The curious case of competition in Spanish speech production. *Language and Cognitive Processes* 21: 760–70.

Vitevitch MS and Stamer MK (2009) The influence of neighborhood density (and neighborhood frequency) in Spanish speech production: A follow-up report. *University of Kansas, Spoken Language Laboratory Technical Report*1: 1–6. Retrieved from: http://hdl.handle.net/1808/5500 (November 2011).

Vitevitch MS, Luce PA, Charles-Luce J, and Kemmerer D (1997) Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech* 40: 47–62.

Appendix 1 Computer-readable characters and the corresponding Spanish phonemes

Spanish ASCII transcription ('Sp-lattese')	IPA		English ASCII transcription ('Klattese')
Stops:			
p	p	*p*elo, *p*en	p
t	t	*t*odo, *t*o	t
k	k	*k*ilo, *k*eep	k
b	b	*b*urro, *v*aso, *b*ut	b
d	d	*d*ía, *d*ay	d
g	g	*g*olpe, *g*o	g
Affricates:			
C	tʃ	*ch*ico, *ch*air	C
–	ʤ	*j*oke	J
Strong fricatives:			
s	s	*s*illa, *s*it	s
–	ʃ	*sh*e	S
–	z	*z*one	z
–	ʒ	a*z*ure	Z
x	x	ca*j*a, relo*j*	–
Weak fricatives:			
f	f	*f*in, *f*ill	f
T	θ	*c*ebra, *c*inta, *z*apato, *th*in	T
(see /b/)	v	*v*an	v
–	ð	*th*en	D
–	h	*h*orn	h
Nasals:			
n	n	*n*ada, *n*od	n
m	m	*m*ono, *m*an	m
–	ŋ	si*ng*	G
3	ɲ	pi*ñ*a	–
Syllabic consonants:			
–	n̩	butt*on*	N
–	m̩	bott*om*	M
–	l̩	bott*le*	L
Glides and semivowels:			
l	l	*l*ado, *l*ate	l
–	ɹ	*r*at	r
r	r	ca*rr*o	–
w	w	p*u*erta, fla*u*ta, *w*in	w
j	i	re*i*na, hac*i*a	–
y	j	pla*y*a, *y*et	y
5	ɾ	ca*r*o	–
6	ʎ	*ll*amar	–

(Continued)

Appendix 1 (Continued)

Spanish ASCII transcription ('Sp-lattese')	IPA		English ASCII transcription ('Klattese')
Vowels:			
i	i	n*i*ño, me*a*t	i
–	ɪ	p*i*n	I
–	ɛ	met	E
e	e	pero, f*a*te	e
–	æ	l*a*ck	@
a	ɑ	n*a*d*a,* lock	a
–	au	l*ou*d	W
–	aɪ	l*i*ke	Y
–	ʌ	l*u*ck	^
–	ɔ	b*ou*ght	c
–	oɪ	b*oy*	O
o	o	nota, note	o
–	ʊ	look	U
u	u	p*u*lgar, c*ou*gar	u
–	ɝ	b*ir*d	R
–	ə	*a*bhor	x
–	ɨ	abandon	|
–	ɚ	butt*er*	X

Notes: The ASCII characters 'x' and 'r' correspond to different computer-readable phonemes in English ('Klattese') and Spanish ('SP-lattese'). If you have English stimulus items and want neighbors that are Spanish words (or vice versa), be sure to replace 'x' and 'r' with characters that are not in either set of computer-readable transcriptions (e.g. '7' or '8') before running those stimulus items through the neighborhood search program for the other language. If a phoneme is found in both languages, a Spanish word and an English word are provided for each phonemic symbol as examples of that sound; the words are not necessarily translation equivalents. Dashes indicate that a sound is not found in the phonemic inventory of that language.