PHONOLOGICAL ROUTE PROCESSING: EVIDENCE OF INTUITION IN PORTUGUESE SPELLING[*]

Fernando César Capovilla[1]
Tiago Tiberio Luz[2]
Luiz Eduardo Graton-Santos[3]
Miriam Damazio[4]

Abstract: A new procedure for calculating spelling difficulty in Brazilian Portuguese is presented. It aims at predicting the risk of committing spelling errors. It is based on phone to grapheme prevalence indexes. For any given phone there are n graphemes, each with its prevalence index. For instance, 11 graphemes encode phone [s], each with its own prevalence: 54% as "s" [sa'livɐ], 27,4% a "c" [si'gahɐ] , 9,2% as "ç" ['pɾasɐ], 5,38% as "ss" ['masɐ], 1,82% as "sc" [fasina'doɾ, 1,51% as "x" [espe'liɾ], 0,40% as "z" ['dɛs], 0,12% as "xc" [ese'deɾ], 0,0079% as "xs" [esu'dʲiɾ], 0,0026% as "sç" [kɾe'sɐ], and 0,0009% as "cç" [se'sɐ̃w̃]. According to it, the lower the prevalence with which a given grapheme encodes a given phone, the greater the spelling error vulnerability. Prevalent graphemes tend to intrude upon non-prevalent ones. A study assessed whether prevalence indexes could account for spelling error distribution. A sample of 154 students (61 college and 93 elementary school ones) was exposed to a spelling under dictation task involving 280 different phone-grapheme prevalence indexes. Each student had to spell 560 rare words, 6.6 phones each on average, totaling 3,676 spelling events. Regression analysis results revealed spelling precision is directly proportional to prevalence index. Thus the procedure has been found empirically valid for calculating spelling difficulty in Brazilian Portuguese.

Keywords: Spelling. Dictation. Phonology. Encoding. Assessment

Resumo: O artigo apresenta novo procedimento para calcular dificuldade de escrita no Português brasileiro, e para prever paragrafias. O procedimento é baseado em índices de prevalência de grafemas por fone, e almeja prever o risco de cometer paragrafias. Esse índice computa a proporção com que cada grafema cifra cada fone na Língua Portuguesa. Para cada fone há n grafemas, cada qual com seu índice de prevalência, expresso em porcentagem. Por exemplo, 11 grafemas cifram o fone [s], cada qual com sua prevalência: 54% como "s" [sa'livɐ], 27,4% como "c" [si'gahɐ], 9,2% como "ç" ['pɾasɐ], 5,38% como "ss" ['masɐ], 1,82% como "sc" [fasina'doɾ, 1,51% como "x" [espe'liɾ], 0,40% como "z" ['dɛs], 0,12% como "xc" [ese'deɾ], 0,0079% como "xs" [esu'dʲiɾ], 0,0026% como "sç" [kɾe'sɐ], and 0,0009% como "cç" [se'sɐ̃w̃]. Quanto maior a prevalência, menor o risco de paragrafia. Relações prevalentes são intrusas nas fracas, mas sofrem baixo risco de intrusão. Um estudo avaliou se índices de prevalência podem prever a distribuição de erros de escrita. Uma amostra de 154 estudantes (61 universitários e 93 de ensino fundamental) foi submetida a uma tarefa de escrita sob ditado envolvendo 280 diferentes índices de prevalência. Cada aluno tinha de cifrar 560 palavras raras, cada qual com 6,6 fones em média, totalizando 3.676 instâncias de

[*] Capes, Observatório da Educação (Obeduc).
[1] Experimental Psychology Departament, Institute of Psychology, University of São Paulo, São Paulo, São Paulo, Brazil. fcapovilla7@gmail.com
[2] Experimental Psychology Departament, Institute of Psychology, University of São Paulo, São Paulo, São Paulo, Brazil. tiagoluz@gmail.com
[3] Experimental Psychology Departament, Institute of Psychology, University of São Paulo, São Paulo, São Paulo, Brazil. luizgraton@gmail.com
[4] Experimental Psychology Departament, Institute of Psychology, University of São Paulo, São Paulo, São Paulo, Brazil. miriamdamazio@gmail.com

codificação. Análise de regressão revelou que a precisão de codificação é diretamente proporcional ao índice de prevalência. Tais dados forneceram validação preliminar do procedimento de cálculo do grau de cifrabilidade do Português brasileiro.

Palavras-chave: Escrita. Ditado. Fonologia. Cifragem. Avaliação

Introductory notes on the use of the terms *phone* and *phonemes*

In Phonetics a *phone* is any distinct speech sound, regardless of whether that exact speech sound is critical to word meaning or not. It is an absolute speech sound that is not specific to any language. It is an unanalyzed speech segment having distinct perceptual properties. It may be a vowel, a consonant, or even a semi-vowel in diphthongs. In contrast, a phoneme is a speech sound of a given specific language that, if exchanged with another phoneme, is capable of turning one word into another word.

The International Phonetic Alphabet (IPA) uses specific symbols to represent phones. According to Lemle (1995), a phonetic transcription, based on phones, uses square brackets ([ ]), whereas a phonemic transcription based on phonemes uses slashes (//).

The spoken word that corresponds to the written word "bico" is made of four phones [ʻb], [i], [k], [ʊ]. That spoken word has the phonetic representation [ʻbikʊ]. Thus, [ʻbikʊ] is the phonetic transcription of the spoken word "bico" using IPA symbols. The written word "bico" is made of four graphemes: "b", "i", "c", "o". Therefore, each phone is represented by a corresponding grapheme. That is the case in nonwords, as well as in words. In this case, when the phone is used as the criterion to distinguish between words in minimal word pairs it is called a phoneme. Thus, the spoken word [ʻpikʊ] differs from the spoken word [ʻbikʊ] by the phoneme /p/. The phones that are the realization of the same phoneme are called allophones. For instance the word "torcer" is the same whether it is pronouced with the *paulistano* accent with its trill variation [toɾ'seɾ], or the rural *caipira* accent with its retroflex variation [toɹ'seɹ], or even with a *fluminense* accent with is aspirated variation [toh'seh]. Thus considering these three pronunciations of the same word, [ɾ], [ɹ], and [h] are all allophones of the same phoneme /R/. In Portuguese, that interchangeability among allophones happens more frequently when the phoneme /R/ occurs on syllabic coda position, rather than on syllabic attack one.

From a cognitive processing standpoint, all phonemes are phones, but not all phones are phonemes. Phones are called phonemes only when they constitute the difference between spoken words in a minimal word pair. That is, when they distinguish between two words, each one with its corresponding different meaning. For children learning to read and write in a synthetic approach, all speech sounds are phones, because the child is not processing word meaning yet. If the child is comparing two spoken words in a minimal word pair, and identifying the speech sound that constitute the difference between them, then the phone works as a phoneme, and the literacy acquition approach is no longer of a synthetic nature but, rather, of an analytical one. However in cognitive processing, the question of meaning as a criterion to distinguish between a phone and a phoneme is somewhat relative. When we present children with a minimal pair of two non-words and ask them to identify what speech segment is different in the second non-word, in that case that speech segment might work as a phoneme, even though the non-word (by definition) has no meaning. For instance, in Portuguese [‘kepʊ] and [‘ketʊ] are spoken non-words. When children are presented with [‘kepʊ], and then with [‘ketʊ], and asked to say in what aspect the latter non-word differs from the former one, they are expected to reply that the difference is the phone [t]. In this case we say that [t] is a phone because there is no meaning involved. However, we could also call /t/ a phoneme because it is a product of an analytic approach of comparing two hollistic pronunciations (whether they constitute words or just non-words) in search for the discrete unit that constitutes the difference between them. Even though very rare (low-frequency) words have meaning whereas non-words have not, for small children, there is basically no distinction between them. So the question of meaning in the present case is irrelevant as a criterion for distinguishing between the concepts of phone and phoneme. Phonemes are always product of an analytic approach of comparing between hollistic pronunciations in search for the unit responsible for their difference, irrespective of the eventual existance of a specific meaning associated with that hollistic pronunciation. That is, irrespective of whether that pronunciation constitute a word or only a non-word. Therefore the present paper employs the term *phone* instead of the term *phoneme*.

The present model

There are basically two approaches to reading and spelling: the single-route approach (SEIDENBERG & MCCLELLAND, 1989) and the dual-route approach (COLTHEART, 2005). The dual-route approach sustains that the process of gaining access to the meaning of print may use either the lexical route or the phonological route (CAPOVILLA, CAPOVILLA, & MACEDO, 2007, 2008; CAPOVILLA et al, 2006; SEABRA & CAPOVILLA, 2010, 2011). The *lexical route* is based on the direct visual recognition of the orthographic form of familiar words. It permits reading words that have exceptional or irregular phone-grapheme correspondences, provided those words have familiar orthographic forms. The *phonological* *route* is traditionally said to operate based on *grapheme-phone conversion rules* that permit constructing word pronunciation. It permits reading both non-words and novel words, that is, words, whose orthographic form is unfamiliar (i.e., whose representation has not been stored in the orthographic lexicon), provided they are made of regular phone-grapheme correspondences. It permits comprehending the meaning of novel words (i.e., having access to the semantic lexicon), provided their decoding results in phonological forms that are familiar to the reader (i.e., whose representation has been stored in the phonological lexicon), so that their utterance sounds familiar to the reader. If the component phone-grapheme correspondences are irregular, their decoding produces unfamiliar phonological forms, which end up not being recognized by the reader. Consequently, in that case the reader is unable to understand the meaning of such irregularly spelled words.

Given that, in the phonological route, access to meaning is necessarily mediated by speech, reading comprehension only occurs when three conditions are met: 1. Words to be read are made of grapheme strings the decoding of which is so regular (i.e., canonical or non exceptional) that it produces precisely the phone sequence of the corresponding spoken word form; 2. Readers have appropriate decoding skills so as to decode in a precise, fluent, and effortless manner thus reading aloud words almost as fluently and naturally as if they were naming the respective objects that the words correspond to; and 3. Readers have a sufficiently developed phonological lexicon, to allow them to recognize their own utterances as corresponding to familiar words.

In the dual-route approach, the writing system is conceived of as a reversible code system that permits encoding speech into print while spelling under dictation, and decoding print into speech while reading aloud. Phonological processing is the most significant cognitive process underlying the development of reading and spelling skills. The phonological recoding involved in such a process is said to be based on conversion rules. In *reading aloud*, the phonological process consists of converting print into speech by means of the application of *grapheme to phone conversion rules* (i.e., *pronunciation rules*). Conversely, the phonological process involved in *spelling under dictation* consists of converting speech into print by means of the application *phone to grapheme conversion rules* (i.e., *spelling rules*).

In order to shed light upon the nature of the phonological processes underlying the phonological route, the present paper proposes a Recoding Model based on recoding intuition (implicit knowledge) instead of on recoding rules (explicit knowledge). Such intuitions are formed by temporal and spatial contiguities and contingencies based on the mere statistical distributions of bi-directional associations between Speech Units and Print Units at the sublexical level (e.g., Grapheme, Digraph, Syllable), as well as at the lexical level (i.e., Word). Speech Units (SU) are arranged in temporal contiguity in speech, and may occur at the phone level or above it (e.g., spoken syllables, spoken morphemes, spoken words). Print Units (PU) are arranged in spatial contiguity in print, and may occur at the grapheme level or above it (e.g., written syllables, written morphemes, written words).

According to the model, in incidental experiences SU and PU may occur associated by temporal and spatial contiguity. In systematic training their co-occurrence may be established, according to if-then contingencies (i.e., if SU then PU). Such an arrangement tends to constitute by induction a gradient of SU-PU associations, which is experienced as encoding intuitions, and expectancies regarding the way SU may be usually encoded when spelled under dictation. Such a systematic arrangement of if SU then PU contingencies during spelling acquisition training constitutes by unconscious induction implicit *knowhow* competence that permits encoding SU sequences in speech into PU strings in print. Encoding rules would be limited to grammar position rules that condition the encoding of a SU into a PU to the position that the SU occupies in the spoken word (i.e., word beginning, word end, in between vowels, preceding a voice or unvoiced consonant). For instance in Portuguese, phone [k] is spelled as "c" in 90,00% of the cases (e.g., [ka'loɾ] -"calor"), as "qu" in 7,83% of the cases (e.g., [ki'taɾ] - "quitar"), as "q" in 1,83% of the cases (e.g., [delĩ'kwẽȷ̃sjɐ] -"delinquência"), as "k" in 0,216% of the cases (e.g., [kiu'i] -" kiwi"), as "ck" in 0,0827% of the cases (e.g., [lamaɾ'kistɐ] - "lamarckista"), as "ch" in 0,314% of the cases (e.g., ['kɾomɐ] -"chroma"), and as "cqu" in 0,0029% of the cases (e.g., [e'kɛɾɪɐ] -"hecquéria"). Thus, upon listening to the phone [k] the expectancy of writing it as "c" is 11,5 times greater than that of spelling it as "qu". However, a position rule permits reducing dependency on mere chance by establishing that "the phone [k] is usually spelled as 'c' when it precedes 'a', 'o' and 'u' (e.g., 'casta', 'costa', 'custa') vowels, and as 'qu' when it precedes 'e' and 'i' vowels (e.g., 'querido', 'quimera')." Drill and practice training develops in a bottom-up fashion the intuitive and unconscious knowhow performance that is primarily basic to spelling competence, whereas grammar rule instruction forms, in a top-down fashion, the formal know-why explanation and decision making patterns that add, secondarily, great efficiency to spelling competence. The present paper argues that, underlying the so called encoding rules are intuitions formed by probabilistic distributions of SU-PU associations, and that the only true rules underlying phonological encoding are grammar rules (i.e., position rules) pertaining to the cases where the PU to be produced depends on the position that the SU occupies in the spoken word to be spelled under auditory dictation by SU-PU encoding.

By the same token, according to the model, in incidental experiences PU and SU may occur associated by temporal and spatial contiguity. In systematic training their co-occurrence may be established, according to if-then contingencies (i.e., if PU then SU). Such an arrangement tends to constitute by induction a gradient of PU-SU associations, which is experienced as decoding intuitions, and expectancies regarding the way PU may be usually uttered when read aloud. Such a systematic arrangement of if PU then SU contingencies during reading acquisition training constitutes implicit *knowhow* competence that permits decoding PU strings in print into SU sequences in speech. Decoding rules would be limited to grammar position rules that condition the decoding of a PU into a SU to the position that the PU occupies in the written word (i.e., word beginning, word end, in between vowels, preceding a voice or unvoiced consonant). For instance in Portuguese, grapheme "s" sounds as [s] 69% of the times, and as [z] 31% of the times. Thus, upon seeing the grapheme "s", the expectancy of sounding it as [s] is 2.2 times greater than that of sounding it as [z]. However, the position rule permits not depending on mere chance and establishes that "the grapheme 's' sounds as [z] when it occurs in between two vowels". Drill and practice training develops in a bottom-up fashion the intuitive and unconscious knowhow performance that is primarily basic to reading competence, whereas grammar rule instruction forms, in a top-down fashion, the formal know-why explanation and decision making patterns that add, secondarily, great efficiency to reading competence. The present paper argues that, underlying the so called decoding rules are intuitions formed by probabilistic distributions of PU-SU associations, and that the only true rules underlying phonological decoding are grammar rules (position rules) pertaining to the cases where the SU to be produced depends on the position that the PU occupies in the written word to be read aloud by PU-SU decoding.

Conversion rules tend to be involved in phonological encoding and decoding only when pronunciation and spelling are position dependent. Yet, even in those cases, pronunciation and spelling may be purely intuitive based on sheer systematic drill and practice exercises centered on instances of PU-SU and SU-PU associations. Even though position rules pertaining to pronunciation rules could describe behavior, those rules may have never been taught or even deduced by the student. It is possible that decoding and encoding intuitions are primary and fundamental to reading and spelling, whereas grammar position rules are secondary and complementary to them. Even though conversion rules can provide *post facto* support for intuition, they cannot substitute for it. The present paper proposes that recoding intuition derives from the statistical distributions of bimodal units (auditory and visual) in bi-directional relationships.

Both reader intuition and writer intuition derive from experience:

(1) The statistical distribution of the relative incidence of all SU that decode a given PU forms Reading intuition. When that PU occurs, the probability to utter one or another of all SU is a positive function of the SU statistical prevalence in that distribution.

(2) The statistical distribution of the relative incidence of all PU that encode a given SU forms Writing intuition. When that SU occurs, the probability to encode one or another of all PU is a positive function of the PU statistical prevalence in that distribution.

PU-SU links are stored in a bimodal (visual-auditory) PU-SU sub-lexicon. Print triggers speech. The Stroop effect (ASSEF, CAPOVILLA, & CAPOVILLA, 2007) results from that encapsulated triggering. Literacy training creates the bimodal Print-Voice sub-lexicon that affords reading, and the bimodal Voice-Print lexicon that affords spelling.

The degree of decoding development and performance derives directly from the statistical distribution of PU-SU links. The degree of encoding development and performance derives directly from the statistical distribution of SU-PU links. The computerized mapping of those distributions (CAPOVILLA & CASADO, 2014) permits predicting both reading error distribution and spelling error distribution.

Natural environments produce overwhelming variation in experience. Experimental and statistical control (via longitudinal and cohort studies) upon children's reading and spelling ontological histories in natural environments are relatively limited. In contrast, to map the distribution of PU-SU links and the distribution of SU-PU links in one's mother language is much more feasible and potentially effective. The present Recoding (Decoding-Encoding) Model uses such a mapping.

Recently a series of papers have proposed a theoretical and experimental model for measuring Portuguese encoding, and estimating the degree of difficulty involved in spelling under auditory dictation any given spoken word in Portuguese (CAPOVILLA, 2011, 2013, 2015a, 2015b; CAPOVILLA & CASADO, 2014; CAPOVILLA & GRATON-SANTOS, 2013; CAPOVILLA, JACOTE, SOUSA-SOUSA, & GRATON-SANTOS, 2011; CAPOVILLA & RAPHAEL, 2004).

Capovilla and Casado (2014) implemented that model in a software named *Brazilian Voice in the New Orthography* (CAPOVILLA, COELHO, & GRATON-SANTOS, in preparation) have the software permits transcribing, in International Phonetic Alphabet (IPA) characters, all different regional pronunciations of any given Portuguese written word. Using the software, Capovilla and Casado (2014) transcribed in IPA characters more that 60,000 words in more than 315,000 different pronunciations. Computerized analysis of the statistical distribution of combinations between SU (in IPA characters) and PU provided a basis for a seminal mapping of the bi-directional relationships between Phonology and Orthography in Portuguese.

The data accruing from such a mapping permit explaining the way phonological route can decode unfamiliar words in reading aloud tasks, and encode those words in spelling under auditory dictation tasks.

The Orthography-Phonology mapping involved two phases:

(1) Identifying each, of all possible SU that may utter any given PU in reading aloud tasks;

(2) Computing the relative percentage of occurrence with which each alternative SU utters any given PU in a representative *corpus* of the Portuguese lexicon. Such a relative percentage corresponds to the Portuguese PU-SU Decoding (GPD) Index, which provides a measure of the degree of decoding difficulty of all PU that comprise Portuguese Orthography.

The present model permits calculating the Word Decoding Degree (WDD) of any given written word to be read aloud. WDD corresponds to the average mean of the GPD indexes involved in pronouncing the written word in a reading aloud task. This involves two steps:(1) Summing up the Decoding (GPD) Indexes of all PU-SU link involved in pronouncing the written word; (2) Dividing that sum by the number of GPD indexes (each for a different PU-SU link) involved in pronouncing that written word.

The Phonology-Orthography mapping also involved two phases:

(1) Identifying each, of all possible PU that may spell (i.e., encode) any given SU while writing in a spelling under dictation task;

(2) Computing the relative percentage of occurrence with which each alternative PU encodes any given SU in a representative *corpus* of the Portuguese lexicon. Such a relative percentage corresponds to the Portuguese SU-PU Encoding Index, which provides a measure of the degree of encoding difficulty of all SU that comprise the Portuguese Phonology.

The present model permits calculating the Word Encoding Degree (WED) of any given spoken word for spelling under dictation. By the way, such a spelling under dictation occurs even in spontaneous writing. In this case, the inner ear responds (via audibilization process, cf. BADDELEY GATHERCOLE, PAPAGNO, 1998) to one's own inner speech. WED corresponds to the average mean of the SU-PU Encoding indexes involved in spelling the spoken word while writing it down under dictation. This involves two steps:(1) summing up the SU-PU Encoding indexes of all SU-PU links involved in spelling the spoken word; (2) dividing that sum by the number of SU-PU Encoding indexes (each for a different SU-PU link) involved in spelling that spoken word.

The software *Brazilian Voice* provides 367 SU-PU Encoding Indexes and their corresponding 367 PU-SU Decoding Indexes. Tables 1 to 3 in the Appendix summarize the SU-PU Encoding Indexes involved in the word list used in the present study. SU-PU Encoding Indexes permit calculating by hand the degree of difficulty of spelling under dictation any given spoken word in Portuguese. PU-SU Decoding Indexes permit calculating by hand the degree of difficulty of reading aloud any given printed word in Portuguese.

Capovilla and Casado (2014) describe an experiment that offers evidence of the model validity for predicting the precision of spelling under dictation in college students. More recently, Capovilla, Coelho, and Graton-Santos (in preparation) have concluded a second version of the software based on cloud computing, and with greatly improved index calculation algorithms based on the analysis of 4,55 million SU-PU link instances. It permits doing without tables, since it calculates automatically the degree of encoding of any spoken word. The present study used such improved estimates, as summarized in Tables 1 to 5 in the Appendix.

The present study

The present study uses the improved model of Portuguese Encoding based on Word Encoding Degrees (WED) and SU-PU Encoding Indexes, and provides a preliminary evaluation of its validity. According to the model, Word Encoding Degree (WED) provides a measure of how easy it is to spell down a word under auditory dictation. The greater the Word Encoding Degree, the easier it is to spell that word under auditory dictation. For any given spoken word, WED corresponds to the arithmetic mean of the SU-PU Encoding Indexes involved in spelling the SU contained in that spoken word. The SU that make up a spoken word have each a given SU-PU Encoding Index. The greater the SU-PU Encoding Index of a given SU, the easier it is to encode that SU in writing (i.e., to spell it down in writing).

In the study, 154 students (61 from college and 93 from elementary school) spelled under dictation 560 very rare spoken words made up of 280 different SU-PU links. In this research, the dependent variable was the total encoding of each word. The total encoding was calculated by summing up the observed percentage of total encoding of each SU-PU link that made up each spoken word and dividing that sum by the number of the SU-PU links present in that word. The independent variable was the encoding degree of each word. The procedure used to establish the encoding degree consists of summing up the SU-PU Encoding Indexes of each SU-PU link that makes up each spoken word, and dividing that sum by the number of

SU-PU links in that word.

In sum: The present study analyzed spelling errors committed by students while spelling under both auditory and visual live dictation. A sample of 154 volunteer students took part of the experiment. There were 61 college students, and 93 elementary-school students from 5th to 9th grade levels. The task consisted of spelling under dictation 560 spoken words of very low orthographic familiarity, so as to maximize phonological encoding processes. The list of 560 words to be spelled under dictation had very low orthographic familiarity, as ascertained by a mean annual search frequency in the Google AdWords database (GROSS et al, 2014) in 16 searches. The task of spelling under dictation consisted in encoding SU into PU. Encoding precision was analyzed as function of the word mean encoding degree, which was the average of the SU-PU Encoding Indexes of all the SU-PU links contained in the spoken word to be spelled. The present study analyzed spelling errors committed under auditory-visual dictation of 560 uncommon spoken words by 154 students (61 college students, and 93 5^{th}-9th grade elementary-school students). Encoding precision was analyzed as function of the word Mean SU-PU Encoding Degree, i.e., the average of the SU-PU Encoding Indexes of all the SU-PU links contained in the spoken word to be spelled under auditory-visual dictation.

Method

Subjects

A sample of 154 volunteers took part in the study. Out of the 154 students, 61 were college students and 93 were elementary school students (18 from 6th grade, 20 from 7th grade, 28 from 8th grade, and 27 from 9th grade).

Instruments

Tables 4 and 5 in the Appendix present a sample of 59 out of the 560 uncommon words for spelling under dictation, along with their precise pronunciation (IPA transcription) and corresponding SU-PU Encoding Degree. The tables list words, along with their respective IPA transcription, SU-PU Encoding Degree, Total Encoding Percentage, and Familiarity Degree (as measured by the Logarithm in the base 10 of the Annual Mean Frequency in the Google AdWords database). The use of relatively novel or rare words aims at reducing lexical recognition and increasing phonological decoding. The use of IPA transcriptions helps standardize the experimenter's pronunciation across sections. During the spelling under dictation task, students used a personal response kit made of 20 sheets of paper containing blank lines numbered from 1-560.

Procedure

Students responded to the spelling under auditory dictation test collectively in their classroom in several 30 min. sessions. On each session, students received their individual response kits. The experimenter instructed students to spell each spoken word in its corresponding line. At each of the 560 items, the experimenter uttered the word preceded by its number. Thus, the experimenter uttered "Word number __ is __" followed by the word. The experimenter repeated the word five times, with a 2s interval in between each repetition. The experimenter dictated words at a rate of 1 per 15-25 seconds. The number of students per session varied. Each student was required to spell 560 words composed of different combinations among 228 types of voice-spelling relationships, in approximately 3,676 instances of voice units to been coded.

Results

There were 521,640 discrete SU-PU encoding opportunities, 285,943 (54.82%) vowel SU, and 235,697 (45.18%) consonant SU. Out of the 521,640 encoding opportunities, 421,126 (80.73%) were appropriate spellings (231,546 vowel SU [80.98%] and 189,580 [80.43%] consonant SU), and 100,514 [19.27%] were misspellings (54,397 [19.02%] vowel SU and 46,117 [19.57%] consonant SU). Out of the 100,514 vowel and consonant misspellings, there were 96,649 [96%] substitutions (51,528 [53%] vowel SU and 45,121 [47%] consonant SU), and 3,865 [4%] were omissions (2,577 pertaining to vowels [67%] and 1,288 [33%] pertaining to consonants). Out of the 96,649 substitutions, 79,518 [82%] were canonical ones (cf. modeled by CAPOVILLA & CASADO, 2014), and 17,131 [18%] were non-canonical ones, which reflect phonetic processes. Out of the 51.528 vowel substitutions, 40,843 [79%] were canonical ones and 10,685 [21%] were canonical ones. Out of the 45,121 consonant substitutions, 38,675 [86%] were canonical ones and 6,446 [14%] were non-canonical ones.

Figure 1 represents total encoding percentage of SU-PU links as a function of the SU-PU Encoding Indexes of the SU-PU links contained in the spelling of those spoken words. Regression analysis of total encoding of SU-PU links as a function of SU-PU Encoding Indexes of those SU-PU links was found significant, F (1, 558) = 3097.96, $p < .0001$, r = .921, r^2 = .847, standardized coefficient Beta = .921, t = 55.66, $p < .0001$.

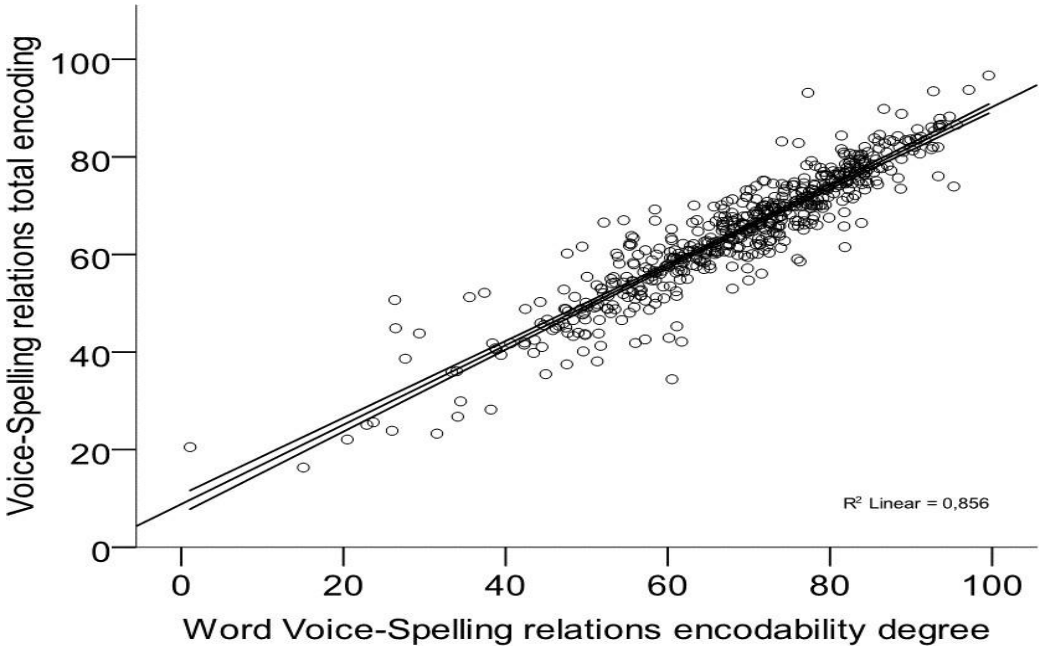

Figure 1. Scatter plot with regression line (confidence interval = .95) of percentage of total encoding of SU-PU links as a function of SU-PU Encoding Indexes of those SU-PU links contained in the spoken words to be written under auditory dictation.

Discussion and Conclusion

The total encoding of SU-PU links in the spoken words to be spelled under dictation increased with the SU-PU Encoding Indexes of the SU-PU links contained in the spoken words to be written under auditory dictation, as computed in accordance with the present SU-PU Encoding Degree Model. The precision of SU-PU Encoding was directly proportional to the magnitude of the SU-PU Encoding Index in Brazilian Portuguese. Over the 521,640 discrete SU-PU Encoding opportunities obtained by the 154 students while spelling the 560 words, it was found that the greater the SU-PU Encoding Index of a given SU-PU link in Brazilian Portuguese, the greater the precision with which a SU was encoded with its corresponding target PU. The greater the dominance of SU-PU links in Brazilian Portuguese, the greater the encoding precision in students spelling under dictation. The greater the recessivity of SU-PU links in Brazilian Portuguese, the greater the incidence of misspellings. The frequency distribution of SU-PU links obtained in the spelling under dictation task was directly proportional to the probability distribution of SU-PU links patterns in Brazilian Portuguese. Spelling intuition (as measured by the probability of encoding a given SU with a given PU) was directly proportional to that probability distribution of SU-PU link patterns in Brazilian Portuguese.

The present study explained how to calculate by hand the Word Encoding Degree of any given Portuguese word. It provided a sample of 228 SU-PU Encoding Indexes out of the

367 SU-PU Encoding Indexes existing in Portuguese. The 228 SU-PU Encoding Indexes presented pertained to the 228 SU-PU Encoding links existing in the list of 560 words to be spelled under auditory dictation in the present task. It provided the Word Mean Encoding Degree of a sample of 57 words out of the list of 560 words to be spelled under dictation in the task of the present study.

Using the 560 word list to assess spelling under dictation in students, the present study demonstrated that the degree of encoding that characterizes each one of the many SU-PU Encoding links in the spoken words to be written under auditory dictation may predict the behavior of choosing from among different PU while encoding.

Results revealed that the probability of choosing a given PU (e.g., grapheme) to encode a given SU (e.g., phone) was a positive function of the magnitude of the SU-PU Encoding Index of the target SU-PU Encoding link involved, as compared with other distracter SU-PU Encoding links. That is, the probability of choosing a target PU to encode a target SU is a positive function of the relative proportion with which that target PU encodes that target SU, as compared to other distracter PU that may encode that same target SU, in a sample of 4.55 million relationships.

Since the SU-PU Encoding Index Model can predict the behavior of choosing from among different PU to encode Portuguese speech, a number of technological and theoretical implications follow. Word Encoding Degree is a scalar variable that permits systematic control in experimental and statistical designs as both, independent variable and covariate. By calculating the Word Encoding Degree of words pertaining to different word lists, one may be able to compare reading and spelling data obtained when using those lists. By doing so, one may contribute to establish productive communication among researchers who use different research instruments. One may also produce word lists free from arbitrary classification of word types, and capable of fine precise tuning. The perspective of regulating systematically word reading difficulty and word spelling difficulty also benefits literacy procedures and instruments at schools, as well as reading and spelling rehabilitation procedures in clinical practice.

In terms of theory, results seem to be compatible with the suggestion that the explanation of the primary mechanisms underlying phonological recoding in reading and spelling may be found on intuition based on statistical distributions of SU-PU links, in addition to conversion rules. If conversion rules are to provide a secondary support for reading and spelling intuition, it is likely that its affectivity depends on the establishment of basal intuitive performance. The greater the intuition, the larger the ground for supporting rules.

Encoding intuition may play an important role in spelling under dictation. It may provide the primary substrate necessary for spelling rules to start making sense and become effective as a secondary support factor. The present study provides support to the idea that encoding intuition results from the statistical distribution of Phone-Grapheme associations in the mother language of literate writers, and that the computerized mapping of that distribution of Phone-Grapheme associations in a representative *corpus* of lexical items can provide Encoding Indexes useful to calculate the Mean Encoding Degree of any given word.

At the first sight the present research data might lead one to the false expectation that the misspelling distribution might be predicted only on the basis of the statistical distribution of phone-grapheme associations in the Brazilian Portuguese. However, that expectation, to use Othero's (2017) expression is but a myth. Othero (2017) lists ten myths regarding language. The tenth myth is that in the future there will exist a universal authomatic translator that shall translate authomatically any phrase from any language into any other language. Indeed, the data yielded by the present paper with regard to non-canonical error frequency pose a limit to the model generality. The present paper showed that out of the 521,640 discrete SU-PU encoding opportunities, there were 421,126 (80.73%) appropriate spellings and 100,514 [19.27%] misspellings. Out of the 100,514 misspellings, there were 96,649 [96%] substitutions and 3,865 [4%] omissions. Out of the 96,649 substitutions, there were 79,518 [82%] canonical ones and 17,131 [18%] non-canonical ones. Out of the 17,131 non-canonical substitutions, 10.685 involved consonants and 6,446 involved vowels.

A detailed examination of the non-canonical errors reveals that they result from phonetic processes that contaminate spelling. Those processes include the following: (1) advancing (misspelling bisbania instead of spelling buxbáumia); (2) anaptyxis or svarabhakti (sinoquisionolona instead of cinoxolona; biliocepetico instead of biliosséptico; abiseço instead of abscesso; flusoquisolou instead of flusoxolol); (3) aphaeresis (siliodes instead of psiliodes; uambóia instead of cuamboia); (4) apocopis (aliasto instead of haliástur; and eliogra instead of heliócrate); (5) desnasalization (siripua instead of siripuã); (6) despalatization (celiado instead of celheado): (7) despirantization (rebaixolite instead of rebaixolice; roco instead of roxo); (8) dissimilation (niocelo instead of lióscelo); (9) epenthesis (fárgea instead of fájea; aguinalhas instead of agnálias; aucoxetando instead of alcochetano); (10) excrescence (eliopetel instead of heliopete; flache instead of flash; kitty instead of kit); (11) gemination (uaade instead of uádi ; aduaanar instead of aduanar); (12) haplology (cecional instead of secessional; asçuapara instead of açuçuapara); (13) hyperthesis, nonadjacent metathesis (alfager ninstead of alfarje; cambio instead of cãibo); (14) lowering (nonileo instead of nonilio); (15) nasalization (unsita instead of hulsita), advancing (helial instead of halial); (16) palatalization (altalha instead of autália); (17) prothesis (reliot instead of heliote; pladexio instead of laudéxio); (18) raising (párcio instead of párseo); (19) simultaneous assimilation and dissimilation (nolilha instead of nonílio); (20) spirantization (amongeava instead of amonjeaba); (21) syncope (equeia instead of hecquéria; and isso instead of hicso; ocilio instead of oxílio); (22) voicing (desencagilhar instead of desencaixilhar; sisiliota instead of siciliota. In a recent paper, Capovilla (2019) proposed a mechanism by which phonetic processes affect writing. According to that author, because novel words have an unfamiliar orthographic form, they are supposed to be written by means of phone-grapheme conversion process in the perilexical writing strategy. Since those words have not only low frequency of occurrence but also low encodability as well, the phone-grapheme encoding tends to be slow. In order to prevent the fading of the phonological word form in the phonological buffer, students tend to reverberate each spoken word form so as to refresh the phonological word represention in the phonological buffer. Each time students reverberate the phonological word form they have heard, phonetic processes intrude in their speech, thus contaminating the phonological form they have heard. Each time the phonological word form that has been heard originally is reverberated, more and more phonetic processes tend to intrude, and thus contaminate, the phonological word form that was originally heard. Eventually, students write down under dictation not what they have heard, but, rather, the phonetic form that they have built while trying to keep on repeating the phonological form they have heard. Phonetic processes are the single most obstacle to the power with which the present model is capable of predicting exactly the error probability at each segment of any spoken word to be written under dictation.

References

ASSEF, E. C. S.; CAPOVILLA, A. G. S.; CAPOVILLA, F. C. Computerized Stroop Test to assess selective attention in children with ADHD. **Spanish Journal of Psychology**, v. 10, p. 33-40, 2007.

BADDELEY, A.; GATHERCOLE, S.; PAPAGNO, C. The phonological loop as a language learning device. **Psychological Review**, v. 105, n. 1, p. 158-173, jan. 1998. CiteSeerX 10.1.1.464.9511. doi:10.1037/0033-295X.105.1.158. PMID 9450375.

CAPOVILLA, F. C. Percepção da fala: como articular propriedades audíveis e visíveis para elevar o Método Fônico a novo patamar. In: A. M. SOARES; J. R. O. R. SIMÃO; L. M. NEVES (Orgs.). **Caminhos da Aprendizagem e inclusão**: entretendo múltiplos saberes. v. 2. Belo Horizonte: Artesã, 2019.

CAPOVILLA, F. C. Paradigma neuropsicolinguístico: Refundação conceitual e metodológica na alfabetização de ouvintes, deficientes auditivos, cegos, surdos e surdocegos. In: F. C. CAPOVILLA (Org.). **Transtornos de aprendizagem**: progressos em avaliação e intervenção preventiva e remediativa. 2 ed. São Paulo: Memnon. 2011. p. 42-131. (ISBN: 978-85- 7954-015-8).

CAPOVILLA, F. C. Paradigma neuropsicolinguístico para refundação conceitual e metodológica da linguagem falada, escrita e de sinais para alfabetização de ouvintes, deficientes auditivos, surdos e surdocegos. In: F. C. CAPOVILLA, W. D. RAPHAEL, & A. C. MAURICIO (Orgs.). **Novo Deit-Libras**: Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Libras) baseado em linguística e neurociências cognitivas. 3. ed. v. 1. São Paulo: Edusp. 2013, p. 73-156. (ISBN: 978-85-314-1330-8).

CAPOVILLA, F. C. Nossa Língua Portuguesa: Como avaliar o grau de cifrabilidade de qualquer palavra falada e o grau de decifrabilidade de qualquer palavra escrita nas normas lusitana e brasileira. In: R. PEREIRA (Org.). **Abordagem multidisciplinar da aprendizagem**. Viseu, Portugal: Editora Qualconsoante. 2015a. p. 481-503.

CAPOVILLA, F. C. Paradigma matricial de linguagem oral, escrita e de sinais: taxonomia e sistema de variáveis para tratamento conceitual, experimental e estatístico. In: F. H. SANTOS; V. ANDRADE; O. BUENO (Orgs). **Neuropsicologia hoje**. 2. ed. Porto Alegre, RS: Artmed Editora. 2015b.

CAPOVILLA, F. C.; CAPOVILLA, A. G. S.; MACEDO, E. C. (Tele)avaliando o desenvolvimento da competência de leitura em ouvintes e surdos de escolas especiais e comuns: O estado da arte. In: E. C. MIOTTO, M. C. S., LUCIA, M. SCAFF, (Orgs.). **Neuropsicologia e as interfaces com as neurociências**. São Paulo, SP: HCUSP. 2007. p. 229-240.

CAPOVILLA, F. C.; CAPOVILLA, A. G. S.; MACEDO, E. C. Analisando as rotas lexical e perilexical na leitura em voz alta: Efeitos de lexicalidade, extensão, regularidade, estrutura silábica e complexidade grafêmica do item e de escolaridade do leitor sobre tempo de reação, duração e segmentação na pronúncia. In: L. PASQUALI (Org.). **Instrumentação psicológica**: fundamentos e práticas. Porto Alegre: ArteMed. 2008. p. 1-24.

CAPOVILLA, F. C.; CASADO, K. **Quebrando o código do português brasileiro**: como calcular a cifrabilidade de qualquer palavra falada e a decifrabilidade de qualquer palavra escrita. São Paulo: Memnon Edições Científicas, 2014. (ISBN: 857954064X)

CAPOVILLA, F. C.; COELHO, R.; GRATON-SANTOS, L. E. **Voz Brasileira na Nova Ortografia**. Versão 2. (in preparation).

CAPOVILLA, F. C.; GRATON-SANTOS, L. E. Compreensão da fala: Relações entre audição e leitura OroFacial, elucidadas por fluxograma de processamento de informação acusticofonêmica e optolalêmica. In: F. C. CAPOVILLA; W. D. RAPHAEL; A. C. MAURICIO (Orgs.). **Novo Deit-Libras**: Dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira (Libras) baseado em Linguística e Neurociências Cognitivas. 3. ed. v. 1: Sinais de A a H. São Paulo, SP: Edusp. 2013. p. 229-233. (ISBN: 978-85-314-1433-6).

CAPOVILLA, F. C.; JACOTE, A.; SOUSA-SOUSA, C. C.; GRATON-SANTOS, L. E. Como calcular o grau de dificuldade na escrita sob ditado das palavras ouvidas: Tabelas de codificabilidade fonografêmica e visibilidade orofacial dos fones do Português brasileiro. In: F. CAPOVILLA (Org.). **Transtornos de aprendizagem**: progressos em avaliação e intervenção preventiva e remediativa. 2. ed. São Paulo: Memnon. 2011. p. 365-391. (ISBN: 978- 85-7954-015-8).

CAPOVILLA, F. C.; MACEDO, E. C.; PENNA, J. S.; CAPOVILLA, A. G. S. Teleavaliação de leitura e escrita em surdos de 5ª. e 6ª. séries incluídos em três escolas públicas comuns do ensino fundamental. In: CONSELHO REGIONAL DE PSICOLOGIA (Org.). **Psicologia e informática**. v. 3. São Paulo: Conselho Regional de Psicologia de São Paulo, 2006, p. 199-234.

CAPOVILLA, F. C.; RAPHAEL, W. D. **Enciclopédia da Língua de Sinais Brasileira**: O mundo do surdo em Libras, Volume 1: Sinais de Libras e o universo da educação, e Como avaliar o desenvolvimento da competência de leitura de palavras (processos de reconhecimento e decodificação) em escolares surdos do Ensino Fundamental ao Médio. São Paulo: Edusp, 2004. p. 1-680. (ISBN: 85-314- 0826-1, ISBN: 85-7060-269-3).

COLTHEART, M. Modeling reading: the dual-route approach. In: M. J. SNOWLING & C. HULME (Eds.). **The science of reading**: a handbook. Oxford: Blackwell. 2005. p. 6-23. http://www.pitt.edu/~perfetti/PDF/Coltheart%2005.pdf

GROSS, M. S.; LIU, N. H.; CONTRERAS, O.; MUÑOZ, R. F.; LEYKIN, Y. Using Google AdWords for international multilingual recruitment to health research websites. **Journal of Medical Internet Research**, v. 16, n. 1, p. 295-305, 2014. http://dx.doi.org/10.2196/jmir.2986. Available at http://www.altgrupo.com.br/adwords/

LEMLE, M. **Guia teórico do alfabetizador**. 10. ed. Rio de Janeiro: Editora Ática, 1995.

OTHERO, G. A. **Mitos da linguagem**. São Paulo, SP: Parábola Editorial. 2017.

SEABRA, A. G.; CAPOVILLA, F. C. **Teste de competência de leitura de palavras e pseudopalavras**: TCLPP. São Paulo: Memnon, 2010. (ISBN: 978-85-7954-011-9).

SEABRA, A. G.; CAPOVILLA, F. C. **Problemas de leitura e escrita**: Como identificar, prevenir e remediar numa abordagem fônica. 6. ed. São Paulo: Memnon Edições Científicas, 2011. (ISBN: 978- 85-85462-98-7).

SEIDENBERG, M. S.; MCCLELLAND, J. L. A distributed, developmental model of word recognition and naming. **Psychological Review**, v. 96, p. 523-568, 1989.

Appendix

Table 1. Voice -Spelling Relationships (VS-Rel) that make up the 560 words and their respective Voice-Spelling Encoding Indexes (VSEI). Part 1: Relationships1-90.

N	VS-Rel	VSEI	N	VS-Rel	VSEI	N	VS-Rel	VSEI
1	[a]-"a"	90.613	27	[aw]-"áu"	1.973	53	[ej]-"êi"	0.211
2	[a]-"á"	9.038	28	[aw]-"ao"	0.345	54	[ej]-"a"	0.164
3	[a]-"ha"	0.333	29	[aw]-"hau"	0.287	55	[ej]-"em"	94.103
4	[a]-"à"	0.001	30	[ɐ̃w̃]-"ão"	99.986	56	[ej]-"en"	5.897
5	[ɐ]-"a"	97.651	31	[b]-"b"	100.000	57	[ej̃]̃ -"ém"	83.133
6	[ɐ]-"â"	2.314	32	[ɔ]-"ó"	66.945	58	[ej̃]̃ -"én"	4.819
7	[ɐ]-"u"	0.015	33	[ɔ]-"o"	32.604	59	[ew]-"eo"	38.289
8	[ɐ]-"hâ"	0.008	34	[ɔj]-"oi"	50.122	60	[ew]-"el"	38.107
9	[ɐ̃]-"an"	80.331	35	[ɔj]-"ói"	49.878	61	[ew]-"eu"	20.225
10	[ɐ̃]-"am"	10.814	36	[ɔw]-"ol"	88.800	62	[ew]-"êu"	2.313
11	[ɐ̃]-"ân"	6.839	37	[ɔw]-"ól"	9.200	63	[ew]-"hel"	0.715
12	[ɐ̃]-"ã"	1.420	38	[d]-"d"	100.000	64	[ɛ]-"e"	53.619
13	[ɐ̃]-"âm"	0.524	39	[dʲ]-"d"	100.000	65	[ɛ]-"é"	40.930
14	[ɐ̃]-"un"	0.015	40	[dʒ]-"d"	99.875	66	[ɛ]-"hé"	0.458
15	[ɐ̃]-"ham"	0.015	41	[dʒ]-"j"	0.028	67	[ɛ]-"a"	0.056
16	[ɐ̃]-"in"	0.005	42	[e]-"e"	97.259	68	[ɛ]-"ê"	0.012
17	[aj]-"ai"	98.285	43	[e]-"ê"	1.386	69	[ɛj]-"ei"	81.986
18	[aj]-"i"	0.591	44	[e]-"he"	1.343	70	[ɛj]-"éi"	18.014
19	[aj]-"ái"	0.355	45	[e]-"é"	0.009	71	[ɛw]-"el"	63.930
20	[aj]-"y"	0.355	46	[e]-"er"	0.001	72	[ɛw]-"éu"	22.388
21	[ɐj]-"ai"	100.000	47	[e]̃ -"en"	85.127	73	[ɛw]-"éo"	5.970
22	[ɐ̃j]̃ -"ãe"	50.000	48	[e]̃ -"em"	7.871	74	[ɛw]-"él"	3.980
23	[ɐ̃j]̃ -"ãi"	5.114	49	[e]̃ -"ên"	6.833	75	[ɛw]-"eo"	1.493
24	[aw]-"al"	67.854	50	[e]̃ -"hen"	0.100	76	[ɛw]-"eu"	0.746
25	[aw]-"au"	26.783	51	[e]̃ -"êm"	0.070	77	[ɛw]-"hél"	0.249
26	[aw]-"ál"	2.432	52	[ej]-"ei"	99.579	78	[f]-"f"	99.992

Table 2. Voice -Spelling Relationships (VS-Rel) that make up the 560 words and their respective Voice-Spelling Encoding Indexes (VSEI). Part 2: Relationships 91-180.

N	VS-Rel	VSEI	N	VS-Rel	VSEI	N	VS-Rel	VSEI
79	[f]-"ph"	0.008	104	[ɪw]-"el"	93.937	129	[k]-"q"	1.829
80	[g]-"g"	94.396	105	[ɪw]-"il"	5.053	130	[k]-"k"	0.216
81	[g]-"gu"	5.555	106	[ɪw]-"hil"	0.040	131	[k]-"ck"	0.083
82	[gz]-"x"	100.000	107	[ja]-"ia"	69.181	132	[k]-"ch"	0.030
83	[h]-"r"	80.550	108	[ja]-"ea"	30.278	133	[k]-"cqu"	0.003
84	[h]-"rr"	19.033	109	[jɐ]-"ia"	95.343	134	[ks]-"x"	91.826
85	[h]-"h"	0.417	110	[jɐ̃]-"ian"	73.485	135	[ks]-"cs"	3.524
86	[i]-"i"	89.300	111	[jɐ̃]-"iam"	14.394	136	[ks]-"cc"	2.643
87	[i]-"í"	10.624	112	[jɐ̃]-"yan"	2.273	137	[l]-"l"	98.689
88	[i]-"y"	0.037	113	[jɐ̃]-"iâm"	1.515	138	[l]-"ll"	1.311
89	[i]-"hi"	0.015	114	[jaw]-"iau"	100.000	139	[m]-"m"	100.000
90	[i]-"hí"	0.013	115	[jɔ]-"ió"	96.591	140	[n]-"n"	99.998
91	[i]-"ie"	0.002	116	[je]-"ie"	98.390	141	[ɲ]-"nh"	99.954
92	[i]-"hy"	0.001	117	[je]-"iê"	1.073	142	[o]-"o"	97.400
93	[ɪ]-"e"	68.546	118	[jɛ]-"ié"	42.857	143	[o]-"ô"	1.872
94	[ɪ]-"i"	28.556	119	[jo]-"io"	76.233	144	[o]-"ho"	0.714
95	[ĩ] -"in"	48.160	120	[jo]-"eo"	23.615	145	[o]-"eau"	0.006
96	[ĩ] -"en"	30.297	121	[jo]-"iô"	0.134	146	[õ]-"on"	81.060
97	[ĩ] -"im"	10.025	122	[ju]-"io"	71.184	147	[õ]-"om"	16.270
98	[ĩ] -"em"	9.166	123	[ju]-"eo"	27.221	148	[õ]-"ôn"	2.107
99	[ĩ] -"ín"	2.068	124	[ju]-"iu"	1.345	149	[õ]-"ôm"	0.340
100	[ĩ] -"ím"	0.163	125	[ju]-"iú"	0.219	150	[oj]-"oi"	100.000
101	[iw]-"il"	88.337	126	[ju]-"yu"	0.010	151	[õj]-"õe"	100.000
102	[iw]-"íl"	4.968	127	[k]-"c"	89.997	152	[ow]-"ou"	61.000
103	[iw]-"iu"	1.728	128	[k]-"qu"	7.842	153	[ow]-"ol"	38.520

Table 3. Voice -Spelling Relationships (VS-Rel) that make up the 560 words and their respective Voice-Spelling Encoding Indexes (VSEI). Part 3: Relationships 181-228.

N	**VS-Rel**	**VSEI**	**N**	**VS-Rel**	**VSEI**	**N**	**VS-Rel**	**VSEI**
154	[ow]-"ôl"	0.120	179	[u]-"ú"	6.288	204	[wɐ]-"ua"	75.325
155	[p]-"p"	100.000	180	[u]-"hu"	0.370	205	[wɐ̃]-"uan"	60.714
156	[ř]-"r"	57.743	181	[u]-"w"	0.163	206	[wɐ̃]-"uam"	23.214
157	[ř]-"rr"	42.257	182	[u]-"oo"	0.035	207	[wɐ̃]-"uã"	5.357
158	[ɹ]-"r"	100.000	183	[u]-"ou"	0.011	208	[waj]-"uai"	100.000
159	[ɾ]-"r"	100.000	184	[ũ]-"un"	53.296	209	[waw]-"uau"	12.000
160	[s]-"s"	54.052	185	[ũ]-"um"	14.072	210	[we]-"ue"	90.000
161	[s]-"c"	27.430	186	[ũ]-"ún"	4.894	211	[we]-"uê"	4.737
162	[s]-"ç"	9.308	187	[ũ]-"úm"	0.482	212	[wẽ] -"uen"	60.965
163	[s]-"ss"	5.372	188	[ũ]-"hom"	0.076	213	[wẽ] -"uên"	39.035
164	[s]-"sc"	1.822	189	[ʊ]-"o"	98.354	214	[wɛ]-"ue"	34.884
165	[s]-"x"	1.492	190	[ʊ]-"u"	1.636	215	[wi]-"ui"	72.662
166	[s]-"z"	0.392	191	[uj]-"ui"	99.008	216	[wi]-"uí"	25.360
167	[s]-"xc"	0.121	192	[uj]-"úi"	0.496	217	[wɪ]-"ue"	81.481
168	[s]-"xs"	0.008	193	[uj]-"ue"	0.331	218	[wĩ] -"uin"	90.291
169	[s]-"sç"	0.003	194	[ũj̃] -"ui"	100.000	219	[wo]-"uo"	97.361
170	[ʃ]-"ch"	10.024	195	[uw]-"ul"	97.434	220	[wo]-"uô"	2.639
171	[ʃ]-"x"	7.382	196	[uw]-"úl"	2.452	221	[ʎ]-"lh"	78.630
172	[ʃ]-"sh"	0.030	197	[uw]-"hul"	0.057	222	[ʎ]-"li"	16.749
173	[t]-"t"	99.974	198	[ʊw]-"ol"	100.000	223	[ʎ]-"le"	4.622
174	[tʲ]-"t"	99.918	199	[v]-"v"	99.883	224	[z]-"s"	73.976
175	[ts]-"zz"	22.222	200	[v]-"w"	0.117	225	[z]-"z"	23.698
176	[tʃ]-"t"	99.893	201	[wa]-"ua"	75.745	226	[z]-"x"	2.280
177	[tʃ]-"tch"	0.038	202	[wa]-"oa"	21.980	227	[ʒ]-"g"	50.672
178	[u]-"u"	93.108	203	[wa]-"uá"	1.922	228	[ʒ]-"j"	18.436

Table 4. Sample of the list of 560 words with respective IPA transcription, Encoding Degree, Total Encoding, and Familiarity (Log10 Annual Mean frequency). Part 1.

Item	Transcript	Word	Encoding Degree	Encoding	Familiarity
1	[visenʊ]	visceno	81.10	79.60	1
2	[asetɛɾju]	ascetério	71.68	72.92	1
3	[omeziʊ]	homezio	68.22	69.48	1
4	[aliastuɾ]	haliástur	68.06	67.18	1
5	[nonilju]	nonílio	79.65	78.41	1
6	[ekɛɾja]	hecquéria	42.29	46.02	1
7	[epizewksɪ]	epizeuxe	70.13	73.61	3
8	[asinasifɔliʊ]	acinacifólio	80.67	78.65	1
9	[paɾsju]	párseo	58.06	57.59	1
10	[aliofɪs]	halíofis	55.66	54.32	1
11	[eɾsɪ]	herse	55.98	55.24	3
12	[iksʊ]	hicso	33.95	34.52	2
13	[omazju]	omáseo	61.53	58.70	1
14	[eliɔkɾatʲɪ]	heliócrate	78.38	76.43	1
15	[dʲiksoniɛjɐ]	dicsonieia	82.39	80.32	1
16	[uwsitɐ]	hulsita	68.21	71.61	1
17	[aliaw]	halial	64.06	61.63	3
18	[kuitaw]	cuintau	71.61	75.38	1
19	[bakuɔw]	bacuol	92.51	90.43	1
20	[askuaɾ]	ascuar	86.40	82.05	1
21	[sejsɪ]	seice	62.40	58.56	3
22	[ebawtʲɪ]	hebaute	59.32	61.40	1
23	[eselĩ]	escelim	61.01	59.74	1
24	[eliɛjɐ]	helieia	73.79	69.99	1
25	[ezɛɾɪzɪ]	exérese	64.51	65.18	4
26	[ewbazɛlja]	eubasélea	64.96	67.21	1
27	[wistɪ]	uíste	61.99	65.50	1
28	[taksonja]	tacsônia	60.85	61.14	1
29	[awziʊ]	ausio	72.11	68.28	2
30	[aʎɛlɐ]	alhela	83.83	83.17	2

Table 5. Sample of the list of 560 words with respective IPA transcription, Encodability, Total Encoding, and Familiarity (Log10 Annual Mean frequency). Part 2.

Item	Transcription	Word	Encoding Degree	Tot Encoding	Familiar
31	[akuometrʊ]	acuômetro	85.69	82.92	1
32	[kuɾanaw]	curanau	83.42	81.71	1
33	[sezɛlju]	sesélio	72.68	70.51	1
34	[ozeaɾ]	ozear	81.80	83.40	1
35	[sipalja]	sipália	70.04	69.29	1
36	[ʃɔkʊɪ]	chócue	47.43	50.71	1
37	[sejsaw]	seiçal	57.72	52.11	1
38	[ɛwsiõ]	hélcion	49.57	41.98	1
39	[asidjo]	ascídio	55.86	50.23	1
40	[aliaɾsju]	haliárcio	56.57	55.99	1
41	[aloksuɾja]	aloxúria	79.15	78.02	1
42	[kuoʒɛlʊ]	cuojelo	78.51	75.64	1
43	[lawdɛksju]	laudéxio	71.58	74.40	1
44	[awtalja]	autália	60.74	62.46	1
45	[omiziʊ]	homizio	66.89	68.65	2
46	[siaw]	siau	56.71	56.89	4
47	[maʎisoɾ]	malhissor	80.19	81.99	1
48	[liɔselʊ]	lióscelo	78.72	76.73	1
49	[sawsju]	sálseo	34.44	32.43	1
50	[kueɾɐ]	cuera	95.60	92.06	3
51	[buksbawmja]	buxbáumia	79.45	78.01	1
52	[tʃiʎɐ]	tílea	53.20	51.12	1
53	[esekaɾja]	excecária	66.12	64.18	1
54	[fuksja]	fúcsia	44.73	45.17	4
55	[paɾɛljo]	parélio	84.41	83.03	3
56	[iɐ̃dõ]	iândom	53.10	56.54	3
57	[kõpaskwo]	compáscuo	65.24	63.41	1

Recebido em março de 2019.
Aprovado em agosto de 2019.