2014-09

PhonItalia: a phonological lexicon for Italian

Goslin, J

http://hdl.handle.net/10026.1/9972

10.3758/s13428-013-0400-8
Behavior Research Methods
Springer Science and Business Media LLC

PhonItalia: a phonological lexicon for Italian

Jeremy Goslin[1], Claudia Galluzzi[2], & Cristina Romani[3]

[1]School of Psychology, University of Plymouth
[2]Fondazione Santa Lucia, i.r.c.s.s., Roma
[3]School of Life and Health Sciences, Aston University

Published in Behavior Research Methods, September 2014, Volume 46, Issue 3, pp. 872-886.

Corresponding Author
Dr Jeremy Goslin
School of Psychology
University of Plymouth
Drake Circus
Plymouth
PL4 8AA
UK
Telephone: +44 1752 584826
Email: Jeremy.Goslin@plymouth.ac.uk

Key words: phonological lexicon, lexical statistics, aphasic errors

ABSTRACT

In this article we present the first open access lexical database that provides phonological representations for 120,000 Italian word-forms. Each of these also include syllable boundaries and stress markings, and a comprehensive range of lexical statistics. Using data derived from this lexicon, we have also generated a set of derived databases, and provided estimates of positional frequency use for Italian phonemes, syllables, syllable onsets and codas, character and phoneme bigrams. These databases are freely available from *phonitalia.org*. This paper describes the methods, content, and summarising statistics for these databases. In a first application of this database, we also demonstrate how the distribution of phonological substitution errors made by Italian aphasic patients is related to phoneme frequency.

INTRODUCTION

Lexical databases are a vital resource for the study of language, providing increasingly comprehensive information on the representations and distributions of words in spoken and written language, as well as behavioural measures of recognition (e.g. Balota et al., 2007). This information plays a fundamental role in the design, control, or interpretation of psycholinguistic experiments, and it is an indispensable component for the modelling of word recognition. As such it could be argued that the development and widespread adoption of these databases has been one of the key supporting factors behind our current understanding of language processing, especially in areas such as lexical access and word recognition.

Lexical databases have been developed for a range of languages, although English is perhaps by far the best served in this respect. Estimates of written word frequency have long been available (Thorndike & Lorge, 1994; Kučera & Francis,1967), and extended with phonological representations in databases such as the MRC Psycholinguistic database (Coltheart, 1981; Wilson, 1988). Additional resources also provide information on ratings of age-of-acquisition or the imageability of words (e.g. Bird, Franklin, & Howard, 2001; Gilhooly & Logie, 1980), and behavioural data, such as reaction times for words in naming and lexical decision tasks (e.g. Balota et al., 2007; Keuleers, Lacey, Rastle, Brysbaert, 2012). Studies in, and of, French and Dutch have also benefited from a rich history and wide coverage of lexical databases (BruLex: Content, Mousty, & Radeau, 1990; BDLex: Pérennou & de Calmes, 1987; de Calmes & Pérennou, 1998; Lexique: New, Pallier, Brysbaert, & Ferrand, 2004; New, Pallier, Ferrand, & Matos, 2001; CELEX: Baayen, Piepenbrock, & van Rijn, 1993), and recent behavioural measures (Ferrand et al., 2010; Keuleers, Diependaele & Brysbaert, 2010). After English, French and Dutch languages, lexical database coverage for other occidental languages becomes relatively sparse, with German described in the CELEX lexicon, and phonological transcriptions and other information available for Spanish (LexEsp: Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000) and Greek (IPLR: Protopapas, Tzakosta, Chalamandaris, & Tsiakoulis, 2012).)

For Italian, we are aware of four freely accessible lexical databases. LEXVAR (Barca, Burani, & Arduino, 2002) provides naming latencies and psycholinguistic variables such as age-of-acquisition, imageability, adult and child frequency measures, and orthographic neighborhood size for 626 simple nouns. Colfis (Laudanna, Thornton, Brown, Burani, & Marconi, 1995; Bertinetto et al., 2005) has estimates of written frequency of use, derived lemmas, and syntactic part-of-speech tags for over 180,000 word forms. Syllables PD/DSS is a database of 2719 orthographic syllables, provided with positional token frequency estimates derived from over 11 million word occurrences. Finally, a database by De Mauro, Mancini, Vedovelli, and Voghera (1993) provides frequency estimates for words across a 500,000 word corpus of spoken Italian. Unfortunately none of these lexica provides phonological transcriptions of Italian words[1], meaning that there is no large-scale database that covers the spoken forms, and associated phonological variables, for this language. It is highly possible that the lack of this type of database stems from the perception that Italian orthography is highly transparent (e.g. Maraschio, 1993), with a relatively simple bi-univocal mapping between grapheme to phoneme that could make word-level phonological transcription largely redundant. However, while Italian can be classified as being towards the extreme end of orthographic transparency, many of the relationships between orthography and phonology are not simple one-to-one mappings. These can require more complex rules that can take account of wider phonological or orthographic contexts (see Burani, Barca, & Ellis, 2006). Moreover, some phonological contrasts are not represented in the orthography, meaning that translation between representations can be a laborious process.

One example of a complex mapping rule relates to velar plosive and affricate sounds, which are both represented in the orthography by ‘g’ and ‘c’ in combination with other characters. The velar plosive /g/ is realized by the letter ‘g’ if followed by the vowels ‘o’, ‘a’ or ‘u’, but by the bigram ‘gh’ if followed by the vowels ‘i’ an ‘e’. In contrast, the affricate /ʤ/ is realized by the letter ‘g’ if followed by the vowels ‘i’ an ‘e’, but by the bigram ‘gi’ if followed by the vowels

[1] Although LEXVAR does provide information on the word initial phoneme of the 626 nouns.

/a,o,u/ (thus, /ge/>‘**ghe**’, /gi/> ‘**ghi’**, /ʤe/> ‘**ge**’ /ʤi/> ‘**gi**’; /go/>’**go**’; /ga/>’**ga**’ /gu/>**’gu**’, but /ʤa/> ‘**gia**’, /ʤo/> ‘**gio**’, /ʤu/> ‘**giu**’). The same rules hold for the unvoiced counterparts of these segments (/k/ and /c/). Some palatal sounds are also represented in the orthography by more than one letter (e.g., fricative /S/>’**sci**’, nasal /N/>**’gn**’, lateral /L/>’**gli**’; but see affricate /Z/>’**z**’). These phonemes, moreover, are always geminated in Italian, but the orthography represents them as a singleton. The Italian phonology has a large number of geminate consonants (e.g., 19 % of consonants by frequency type are geminate) and germination is a contrastive feature for the majority of consonants (e.g., pala [spade] vs. palla [ball]; poro [pore] vs. porro [leek]). The phonemes listed above, however, are present *only* in geminate form. Therefore, the orthography does not represent what would amount to redundant information (e.g azione> az.zjo.ne; agnello>aN.Nel.lo, aglio>aL.Lo). Another example is the grapheme ‘h’, which has no phonological counterpart but is still contrastive in orthography (e.g., hanno [They have] vs. anno [year]). Conversely, the phonological contrast in openness between /e/ and /ɛ/ and /o/ and /ɔ/ in standard Italian[2] can be lexically distinctive (e.g., /pɛska/ [peach] vs. /peska/ [fishing]) in stressed syllables, but these phoneme pairs are represented by the single graphemes ‘e’ and ‘o’ respectively. While stress can provide a cue to vowel aperture, with ‘e’ or ‘o’ usually corresponding to open vowels in stressed syllables (e.g. fra.tɛl.lo [brother] and fɔ.to [photo]), the frequent exceptions (e.g. in.so’r.ge.re [to rebel]) mean that this cue is indicative at best, requiring that phonological vowel aperture is established on an item by item basis.

The types of irregularities described above mean that Italian orthography does not provide a sufficiently accurate representation of the Italian phonology for many applications, from robust control of psycholinguistic stimuli, to statistical examinations of cross-linguistic contrast, to analyses of frequency effects in children and in language impaired populations (e.g., aphasic patients, children with specific language impairments). In this paper we

[2] This phonological contrast is also present in some Italian varieties (such as Roman). In others the opposition in vowel height could be neutralized, conditioned by phonotactic factors, or even result in a different lexical contrast (Maturi, 2009).

present an open access lexical database designed to fill this gap, by providing phonological transcriptions across a wide range of Italian word-forms as well as a range of derived psycholinguistic variables, such as phonological neighborhood measures, plus statistical summaries of phoneme and syllable use. This paper describes the methodology behind the construction of this database, describes the information provided in the lexical and derived databases, and provides statistical summaries of the data held within them. We will also present an example of the usefulness to this database by applying a study designed to examine aphasic's phonological errors. Another example of how the statistics derived by the database can be used to inform our understanding of language processing and its universal basis is presented in (Romani, Galluzzi, & Goslin, submitted).

METHODOLOGY

The basis for this lexicon was Colfis (Laudanna et al., 1995; Bertinetto et al., 2005), a database of written Italian word forms derived from 3,798,275 textual occurrences from a corpus of newspapers (1,836,119), magazines (1,306,653), and books (655,503) published between 1992 and 1994. This originally consisted of 188,792 word-forms each with fields describing their part-of-speech tag, and the frequency of occurrence across the three textual sources. Using these Colfis word-forms, we made an initial screening to remove all entries that contained non-alphabetic characters apart from the apostrophe. This resulted in the removal of 44,376 phrases (such as “in giro”) and 1,266 non-words (such as “-se-“), and minor corrections to 2,294 word-forms (for example, changing the entry “canaletto (m)” to “canaletto”). The remaining word-forms were then subjected to further manual screening, resulting in the removal of an additional 5,939 non-words (such as “fndo”) and 17,211 imported words (such as “Dorothy”). It should be noted that not all imported words were removed in this screening process, any considered to be in current usage (such as ‘film’ or ‘Marx’) remain in the database.

At the end of the screening process exactly 120,000 word-forms remained (63.56% of the original Colfis word-forms) as candidates for phonological transcription. The first stage of the process was implemented using the phonological transcription module from the

Italian Festival text to speech system (Cosi, Gretter, & Tesser, 2000). This generated a phoneme string for each of the word-forms with additional markers for syllable boundaries and primary syllable stress. These representations were then converted from Festival's SAMPA phonemic alphabet to a custom alphabet in which each of the 29 individual Italian phonemes labelled in the lexicon could be presented by a single standard text character, as described later in Table 2. It is worth noting that this transcription does not make a distinction between the alveo-palatal fricatives /s/ and /z/. This is because these phonemes are not used in a contrastive fashion in Italian, and differences in their distribution are a matter regional preferences or an allophonic variation dependent on context. For example, the unvoiced allophone /s/ is used before voiceless consonants (as in 'scarpa') while the voiced allophone /z/ is used before voiced consonants (as in 'sgravio'). Since our aim was to provide a phonological and not a phonetic description of Italian words, we transcribed both allophones with the same symbol (/s/; see later sections for more details). The placement of syllable boundaries was then modified where necessary to conform to Italian-specific syllabification rules based upon those created by Laporte (1993) for French. These rules dictate minimal syllable onsets; such that the syllable boundary should be placed before the last segment of an intervocalic consonant cluster which is not a glide (see Goslin & Frauenfelder, 2001 for a comparison of syllabification algorithms). This means that intervocalic syllable onsets would consist of a single consonant by default, such as in /vOl.ta/ ('volta'), /as.ta/ (pole)*. Exceptions, however, involve obstruent segments which are immediately followed by a liquid (e.g. /pl/) since these clusters are treated as tautosyllabic. Moreover, if there is a glide immediately preceding the vowel then the onset is extended to include another consonant, if one is available, such as in /stO.**rj**a/ ('storia') or /GraZ.Zje/ ('grazie'). Finally, both exceptions can combine to produce an onset consisting of an obstruent, liquid, and glide, such as in /is.trja/ ('istria').

Each of the generated phonological representations (and syllable stress and boundary markers) was then manually checked by the second author, with additional random spot-checking from the final author, both of whom are Italian native speakers. Any disagreements were settled by discussion. The transcription was intended to conform to a standard Italian pronunciation that is generally uncontroversial, apart from some alternations between /e/-/ɛ/ and /o/-/ɔ/, which are subject to regional variations. Even in these cases representations are intended to approximate a 'standard' pronunciation, although both of these native Italian linguists have the regional accent of Rome which may colour their judgements. Multiple redundant checking meant that each phonological representation was verified at least twice. It was found that 28,168 representations required some form of manual correction (30.67% of the lexicon).

An evaluation of the reliability of the phonological representations was made via blind phonemic transcription of 500 word forms selected at random from the database. These were hand transcribed using the phonetic alphabet adopted by *phonItalia* by a native Italian speaker that was independent of the development of the lexicon. Point-to-point agreement was calculated between each of the 2917 phonemes representing those 500 words in the database and the independent transcription. Phonemic insertions or deletions made by the independent transcriber not found in the lexicon were also counted as errors. This comparison revealed phonemic agreement of 98.35%, with a Kappa of 0.983. It should be noted that the majority of the disagreements (28 of a total of 48) were due to differences in the marking of vowel aperture (/e/-/ɛ/ or /o/-/ɔ/); likely due to regional differences in the representations used by the original *phonItalia* linguists (Rome) and that of the independent transcriber (Florence).

LEXICAL STATISTICS

As described in the previous section, this new lexicon provides phonological representations for 120,000 Italian word forms, along with associated syllable boundary and stress markers. While the Colfis database provides frequency, part-of-speech tags and the lemma for each word-form (a description of original *Colfis* fields is provided on *phonItalia.org*), *phonItalia* augments this information with a range of additional fields that provide information related to both the phonological and orthographic representations of the words.

Additional orthographic fields include the consonant vowel structure of the word, the number of homographs of that word, and the uniqueness point, that is the letter *at which* the orthographic representation becomes unique. As the uniqueness point lists a value of zero if the representation never becomes unique, an additional field is also included which lists the uniqueness point minus one (*OrthUniqM1*). For non-unique words this field will have a value of the length of the word, and thus avoids the potential skewing in summarising statistics that could result from the zero values of the uniqueness point field. All of these fields have also been reproduced for the phonological representation of the words, with a number of further additions. For the phonological vowel consonant structure, consonants that are in geminate pairs are given the representation 'G' rather than 'C'. For example, /kap.pot.to/ is /CVG.GVG.GV/. Other fields have been added that relate to syllabic information, listing the number of syllables in the word, the position of the stressed syllable, and a phonological representation that includes syllable boundary markers (denoted by '.').

Each word is also provided with estimations of both orthographic and phonological neighborhood, these have been estimated using measures of Colheart's N (Coltheart, Davelaar, Jonasson, & Besner, 1977) and Levenshtein distance. Coltheart's N is calculated as the number of lexical character sequences that can be constructed by changing a single character of the current entry while the position and identity of the remaining characters remain unchanged. All neighboring lexical entries that are homographs were grouped and counted as a single neighbour. The Levenshtein distance is the number of single insertions, deletions, or substitutions required to change from one character string to another. To calculate this value the Levenshtein distance between the orthographic representation of the current entry and all other unique orthographic/phonological entries in the lexicon are calculated. The reported orthographic/phonological Levenshtein distance (OLD/PLD) 20 being the mean of the 20 smallest distances found. Additional fields related to these metrics include estimates of the total frequency of neighbors, and also estimates of the number and frequency of those with higher or lower frequency than the target word. Finally, the main *phonItalia* database also provides mean and summed frequencies of the orthographic and phonological bigrams contained within each word (individual character-bigram and biphone statistics are also made available in a separate derived database described below).

All fields that required calculation based upon estimate of frequency of use (such as *Phon_N_MFreq,* mean log[3] frequency of words in the phonological neighborhood), we based this upon the *Colfis* total frequency estimate field *fqTot*. All of the new data fields included in *phonItalia* are shown in Table 1, along with a summary of the global statistics for numeric fields calculated across the entire lexicon.

DERIVED SUB-LEXICAL STATISTICS

The provision of phonological word forms within this lexicon allows for the first comprehensive estimation of the relative frequency of occurrence of Italian phones, syllables, and other phonological representations. These have been calculated across all word forms within the lexicon to produce both non-positional and positional type and token frequency measures. Type frequency measures (identified by the fields *TypeF)* refer to the number of times a particular unit (phoneme, syllable, etc.) occurs within the words of lexicon with each word counted once. Token frequency (identified by the fields *TokenF,* with the natural log of this value found in the field *LnTokenF)* refers to the number of times a unit occurs in the words of the language taking into account the frequency of the words. Thus, phoneme occurrences are multiplied by the frequency of the words in which they occur and then summed. All token frequencies are calculated using total lexical frequency measure from Colfis (field name *fqTot*). Multiple instances of a unit within a word are additive, so the type count for /p/ would be incremented twice for the word /prO.prjo/ (‘proprio’), and the token count increased by twice its lexical frequency (2 * 2408). Estimates for phone frequency are provided both overall and relative to syllabic position (see below for more details). In addition, overall frequency data for different types of multi consonantal syllable onsets are provided (e.g. the frequency of onsets like, /p/, /pr/ ,/pl/ or /str/). Syllable frequencies are

[3] All log frequencies are calculated using the natural log.

provided overall and according to word position. Character-bigram and biphone frequency statistics have also been calculated across the lexicon, with frequency estimates provided relative to word and (for biphones) syllable position. This information is provided in a number of additional databases separate to the main lexicon, the contents of which are summarised in the following sections. As with the main lexicon, all these additional databases are available from the lexicon website in Excel, and tab-delimited text format. The source code to and program used to generate these derived statistics (as well as update statistics in the main word forms database – such as bigram frequency or uniqueness points) are also available in from the database website, *phonItalia.org*.

Phone Statistics

This database provides the frequency of occurrence for all 29 Italian phones used within this lexicon. Overall phonemic frequency of use are summarised in Table 2, with the database also providing statistics for phones relative to specific syllabic positions. These fields are as follows:

Single Onset provides statistics for phones found in a single consonant syllable onset. For example, the phone /n/ in the word /a.E.ro.pla.**n**o/.

Onset /Cc/ for phones found in the first consonant of a double consonant syllable onset. For example, /p/ in /a.E.ro.**p**la.no/.

Onset /cC/ for phones in the second consonant of a double consonant syllable onset. For example, /l/ in /a.E.ro.p**l**a.no/.

Onset /Ccc/ for phones in the first consonant of a triple consonant syllable onset. For example, /G/ in /Gan.**G**ljo /.

Onset /cCc/ for phones in the second consonant of a triple consonant syllable onset. For example, /l/ in /Gan.G**l**jo/.

Onset /ccC/ for phones in the third consonant of a triple consonant syllable onset. For example, /j/ in /Gan.Gl**j**o/.

Nucleus for phones that form the nucleus of a syllable. For example /o/ is twice found as a nucelus in /a.E.r**o**.pla.n**o**/.

Single Coda provides statistics for phones found in a single consonant syllable coda. For example, /n/ in the word /la**n**.ce/.

1st Coda for phones in the first consonant of a syllable coda (greater than one consonant in length). For example, /l/ in /fi**l**m fi**l**m/.

2nd Coda for phones in the second consonant of a syllable coda (greater than one consonant in length). For example, /m/ in /fil**m**/. There are very few of these cases in Italian.

Geminate provides statistics on phones that are found in geminate position in a word. For example, /g/ in the word /ma**g**.**g**o.re/. Table 3 provides a summary of the relative frequency of consonant occurrence when geminate (e.g. /n/ in /dO**n**.**n**a/ ‘donna’) or non-geminate (e.g. /n/ in /pu**n**.to/ ‘punto’).

Syllable Statistics

This database contains calculations of the frequency of use for the 3626 unique syllables found within the lexicon. An observation worth noting is that phonological syllables appear to be far more numerous[4] (33% more types) in Italian than orthographic syllables, with only 2719 listed in PD/DPSS Syllables (Stella & Job, 2001). This serves to highlight the degree of ambiguity between the Italian orthography and phonological representations. A summary of the distribution of phonological syllabic frequency by syllable length is shown in Table 4, with a similar summary of syllable stress as a factor of length in Table 5. As in the phone database type and token frequencies are provided for all occurrences, irrespective of their word position, with additional statistics for occurrences in specific word position, as follows:

[4] Despite PD/DPSS Syllables being based upon a corpus of 143,970 word types verses the 120,000 in phonItalia.

MonoSyll provides frequency information for syllables that occur in monosyllabic words.

Initial is the field that describes syllables that occur word initially in multisyllabic words, for example /ti/ in /ti.fa.no/.

Medial provides frequency information for syllables from multisyllabic words that are not in either word initial or word final position, for example /ti/ in /ul.ti.mo/.

Final gives frequency information for syllables found in multisyllabic words that are word final, for example, /ti/ in /van.ti/.

A subset of this syllabic frequency information, containing the 100 most frequent syllables is listed in Appendix A, ordered by token frequency. In addition to the overall syllabic data, each syllable in the database is also provided with additional fields with the frequency of occurrence for the corresponding phone sequence irrespective of syllable boundaries. The previous syllable fields only include frequencies for phone sequences that respected syllable boundaries, such as the syllable /par/ in the word /**par**.ti.ta/. In the following n-Gram type sequence frequency statistics, the token and frequency calculations also include occurrences of the same phone sequence that cross syllable boundaries, such as /par/ in the word /pre.**pa**.**r**a/.

PhonSeq_Total gives the frequency of occurrence for the phone sequence of the syllable in the lexicon irrespective of syllable boundaries.

PhonSeq_Word_Initial is similar to *PhonSeq_Total* but only includes the statistics for words where the syllable phone sequence is found word initially. For example, statistics for the syllable /tar/ would include an occurrence for the word /**ta**.**r**a.re/, but not in /kon.**ta**.**r**e/.

Syllable Onsets and Codas

To complement the previously described syllabary, separate databases are also made available that describe each of the 132 syllable onsets and 58 syllable codas, summarised by length in Table 6. In these databases, the type and token frequencies of each particular onset or coda are provided. The onset and coda databases also list a blank entry that has been included to provide statistics for the occurrence of syllables with an empty onset (e.g. the syllable /ar/) or coda (e.g. the syllable /si/). As in the syllabary, these statistics are provided for all occurrences irrespective of word position, plus those found in particular word position, as described below.

Total gives statistics for syllable onsets or codas found in any word position

Word Initial gives statistics for syllable onsets found in word initial position, for example, /t/ in /**t**i.fa.no/

Word Medial provides statistics is provided for both syllable onsets and codas that are medial to the word. For example, the onset /d/ or the coda /n/ in /mon.**d**o/

Word Final provides statistics is provided only for syllable codas that are found in word final position.

Geminate is a subset of the word medial statistics, and is limited to syllable onsets or codas that are geminate, for example, the onset and coda /l/ in /a**l**.**l**o/.

For clarity, syllable onsets and codas have also been split into their constituent consonants, with each consonant held in separate fields.

Number of phones is the number of phones in the syllable onset or coda.

1st phone is the 1st (leftmost) phone in the syllable onset or coda, for example /p/ in the onset /**p**l/, or /l/ in the coda /**l**m/.

2nd phone is the 2nd phone in syllable onset or coda, for example /l/ in the onset /p**l**/, or /m/ in the coda /l**m**/.

3rd phone is the 3rd phone in syllable onset or coda, this would be blank in the example of /pl/, or would be /rs/ in the coda /rk**s**/ from ‘Marx’.

4th phone is the 4th phone in syllable onset (this field is missing in the coda database).

Character-bigram and Biphone Statistics

Two separate databases provide statistics covering 577 biphones and 478 character-bigrams calculated across the lexicon. This information is provided for all occurrences, but additional statistics are provided for occurrences relative to word position, with biphones also having statistics for occurrences relative to syllable position.

Word Initial gives the statistics of bigrams that occur in word initial position. For example, the biphone /ko/ in /**ko**n.trad.det.te/ or the character bigram ‘se’ in ‘**se**mpre’.

Word Medial has statistics for bigrams that occur word medially, For example, the biphone /**on**/ in /k**on**.trad.det.te/ or the character bigram ‘mp’ in ‘se**mp**re’.

Word Final gives frequency information for bigrams that occur word finally. For example, the biphone /te/ in /kon.trad.det.**te**/ or the character bigram ‘re’ in ‘semp**re**’.

Syllable Onset gives frequency statistics for biphones that are found in syllable initial position, for example /tr/ in /kon.**tr**ad.det.te/. This would include all occurrences in which the first and second phone of the biphone and syllable were shared.

Syllable Medial provides statistics for biphones found in syllable medial position, for example /ra/ in /kon.t**ra**d.det.te/. This would include all occurrences where neither the first or second phone of the biphone coincided with the initial or final phone of a syllable.

Syllable Final gives frequency statistics for biphones that are found in syllable final position, for example /et/ in /kon.trad.d**et**.te/. This would include all occurrences in which the final and penultimate phone of the bigram and a syllable were shared.

Cross Syllable biphones are those that cross syllable boundaries. For example, /nt/ in /ko**n**.**t**rad.det.te/. In this case the first phone of the biphone must consist of the final phone of the syllable preceding the boundary, and the second phone the first phone of the syllable that proceeds the boundary.

Orthographic Character Statistics

This database contains calculations of the frequency of use for 27 orthographic characters used in the word forms of the lexicon, including the apostrophe, irrespective of word position.

APPLICATION OF LEXICAL STATISTICS TO ANALYSES OF APHASIC ERRORS

Analyses of speech errors have played a very important role in constraining models of speech production, and they are a crucial tool to diagnose the level of impairment in patients suffering from language difficulties following a stroke (aphasia). While analyses of the relationships between word frequency and errors are routinely used as a diagnostic tool, analyses of the influence of phoneme frequency have been very limited in their scope.

Early studies by Blumstein (1973;1978) found no difference in frequency effects between small groups (n ~= 6) of Broca, Wernicke's and Conduction aphasics However, a larger study by MacNeilage (1982) contrasted 20 English-speaking non-fluent aphasics (with possible apraxic difficulties) with 10 fluent aphasics. He found that target error rates were greater in low than high frequency phonemes (frequency correlated with % of errors), but only in the non-fluent group. In contrast, the incidence of intruding segments was found to increase with phoneme frequency across both groups, an effect also found by Robson, Pring, Marshal and Chiat (2003) in a fluent patient with jargon aphasia. Goldrick and Rapp (2007) also reported contrastive effects, with an effect of frequency in a patient with a post-lexical locus, but not in a patient with lexical phonological impairment.

An examination of the limited evidence from these studies suggest that it may only be apraxic patients, those with articulatory difficulties, who have greater difficulties in computing the articulatory programs associated with low frequency phonemes. This hypothesis would predict an inverse relationship between articulatory complexity and phoneme frequency, with high frequency phonemes being easier to articulate. For other patients, phonological errors do not appear to be due to difficulties in computing articulatory programs, but they occur because of confusion in lexical representations or difficulties in selecting the right phonemes for a word. For these patients, frequency will not affect the ability to produce target phonemes, although more frequent phonemes may still be selected erroneously over the actual targets.

In our study we examine whether the relationship between phoneme frequency measures from *phonItalia* and the distribution of production impairments can be used to distinguish between these different of types of aphasic patients.

Method

Patients: Two patient sub-groups were selected from a patient pool of 24 patients, all of whom had confirmed diagnosis of aphasia. Of these 22 had suffered from left hemisphere stroke, one from right CVA, and one from close head injury. All had been selected due to the high number of phonological errors they exhibited across a range of speech production tasks, an absence of peripheral dysartric difficulties (e.g., systematically distorted speech), and relatively good phonological discrimination abilities. Further details of this particular set of patients can be found from previous studies (see Romani, Galluzzi, Bureca & Olson, 2011; Romani, Galluzzi & Olson, 2011, and also Romani & Galluzzi, 2005). Subgroups were selected on the basis of particularly high or low rates of phonetic errors. The 11 members of the *phonological-apraxic (*ph-apraxic) group were selected because they made more than 10% of phonetic errors, while the nine *phonological-selection* (ph-selection) patients made fewer than 5% phonetic errors.

Task and Analyses: Patients were asked to repeat 773 words, with a phonemic transcription made of their repetitions. Analyses were limited to phoneme substitution errors. Following the procedure of MacNeilage (1982), we examined the correlation between the percentages of times a phoneme was substituted in error (replaced rates) and its token frequency from *phonItalia*. We also conducted a separate analysis of the correlation between the number of times each phoneme type was used instead of targets in the substitution errors (replacing numbers) and its token frequency count. Phonemes /N/, /L/, /S/ and /z/ were removed from the analyses as these segments are always geminate, which could have reduced error rates. Deletion and insertion errors were not included in the analyses. Patients generally avoid the production of phonotactically illegal sequences and/or difficult sequences of vowels and, for this reason, only a limited set of consonants can be deleted (sonorants in certain syllabic positions ,see Romani et al., 2011a, for an explanation). This limits the potential scope of analyses on deletion and insertion errors.

Results and Discussion

A summary of the results can be seen in Table 7. It was found that there was a significant negative correlation between the percentage of substitution errors and phoneme frequency in the ph-apraxic patients ($r = -0.50$, $p < 0.05$), but no significant correlation in the ph-selection patients ($r= -0.22$, $p=0.36$). An examination of the relationship between the number of times a phoneme was used as a replacement and its frequency revealed significant positive correlations in both the ph-apraxic ($r=0.55$, $p < 0.05$) and the ph-selection ($r=0.87$, $p < 0.001$) patient group. We also conducted linear regression analyses with frequency and patient group as predictors of rate of errors on the different phonemes and number of times different phonemes were used as replacements. For rate of errors, we found a marginally significant interaction between frequency and group ($F(1,33)=3.93$; $p=.056$). Individual analyses showed that frequency was significant for the apraxic groups ($F(1,17)=5.26$; $p=.036$), but not for the phonological group ($F(1,17)=0.85$; $p=.37$). The linear regression predicting the number of times different phonemes were used as replacements showed no significant interaction between frequency and group ($F(1,33)=2.01$; $p=.17$), but there was a significant main effect of frequency ($F(1,34)=13.6$; $p<.001$).

The error rates results support our original diagnostic division between patients where phonological errors are motivated either by difficulties with the articulatory production of the phonemes (in the ph-apraxic group), or by difficulties in the selection of the right phonemes (in the ph-selection group). Moreover, it also points towards a relationship between phoneme frequency and articulatory complexity. Frequency influenced rate of substitutions only in the ph-apraxic group. It is possible that, in this group, errors on the low frequency segments are more likely because generally these are the segments more difficult to articulate. These results are consistent with those of an earlier study (MacNeilage, 1982), and also with findings of the effects of syllable frequency in patients with apraxia of speech (Aichert & Ziegler, 2004; Steiger & Ziegler, 2008), but not in patients with more central phonological difficulties (Wilshire & Nespolous, 2003; but see also Laganaro, 2008 for inconsistent results). These findings lend support to studies showing how phonological complexity and frequency can be used to *selectively* identify characterizeapraxic patients (Romani & Galluzzi, 2005; Romani, Granà, & Semenza, 2002; Romani et al., 2011a). Both analyses of frequency and complexity highlight important differences between types of patients that are not well recognised in the literature, but that can have important implications for diagnosis and rehabilitation (see Blumstein, 1973; 1978).

Our results also revealed a significant positive correlation between the frequency of phonemes and how many times they are used as replacing phonemes across both patient groups. This result is an apparent contrast with the results of a recent study where we show that articulatory complexity does not influence which phonemes are used as replacement in the phonological group (Galluzzi, Bureca & Romani, submitted). It is possible, however, that, although strongly related, frequency and articulatory complexity of phonemes are partially independent variables. Thus, for patients *without* articulatory difficulties, frequency is a stronger variable than complexity in informing choice among alternatives and, therefore, in determining which phonemes are used as replacements. Similarly, in Romani, Galluzzi and Goslin (submitted), we found that complexity and frequency were strongly correlated when predicting age of acquisition in Italian children, indicating that within-language phoneme frequency is influenced by articulatory complexity. However, it must be noted that data from the latter study also point to other factors, independent of complexity, that influence the distribution of phoneme frequency.

CONCLUSION

The primary aim of this project was to produce a lexical database for Italian that would include the phonological transcriptions of word-forms. This database includes a comprehensive set of common psycholinguistic variables to cover both the spoken and written modality. The first use of this resource has been to produce a set of derived databases that include frequency of use statistics for Italian across a range of units, including both phonemic and syllabic units. All of these databases are open access, available from the website *phonItalia.org* formatted in Excel, and tab-separated text format, freely distributed under a creative commons license. This resource will be of utility across a wide range of research, from the design or analysis of psycholinguistic experiments with Italian stimuli, natural language processing, and in cross-linguistic applications. It is hoped that the distribution of this database under an open access license will encourage further extensions or changes to the databases in the future. Finally, we have shown how important conclusions can be derived from applications of some our derived statistics. In particular, we demonstrated that analyses of phoneme frequency (as well as word frequency) on speech errors can provide important cues to the locus of an individual's language impairment.

REFERENCES

Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, *88*(1), 148-159.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The Celex lexical database (CD-ROM).* University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445-459.

Barca, L., Burani, C., & Arduino, L.S. (2003). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, and Computers, 34 (3),* 424-4Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., & Thornton A. M. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. http://linguistica.sns.it/CoLFIS/CoLFIS_home.htm

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words , *Behavior Research Methods, Instruments, and Computers, 33,* 73-79.

Blumstein, S. (1973). Some phonological implications of aphasic speech. In H. Goodglass and S. Blumstein (eds.), *Psycholinguistics and Aphasia*. Baltimore: Johns Hopkins University Press, pp. 123-236.

Blumstein, S.E. (1978). Segment structure and the syllable in aphasia. In A. Bell and J.B. Hooper (Eds.), *Syllables and Segments*, pp.189-200. Holland: North-Holland Pub. Co.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Burani, C., Barca, L., & Ellis, A.W. (2006). Orthographic complexity and word naming in Italian: Some words are more transparent than others*. Psychonomic Bulletin and Review, 13,* 346-352.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies ased on film subtitles. *PLOS ONE, 5*, e10729.

de Calmès, M., & Pérennou, G. (1998). BDLEX : a Lexicon for Spoken and Written French. *In: 1st International Conference on Langage Resources & Evaluation (LREC1998), Grenade. ELRA, Paris*, p.1129-1136, 28-30 mai 1998.

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica, 32*, 133-143.

De Mauro, T., Mancini, F., Vedovelli,M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato [frequency lexicon of spoken Italian]*. Milan: ESTALIBRI.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Cosi, P., Gretter, R., & Tesser, F. (2000). Festival parla italiano, in *Proceedings of GFS2000, XI Giornate del Gruppo di Fonetica Sperimentale*, Padova, 29th November to 1st December.

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

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488-496.

Gilhooly, K.J., & Logie, R.H. (1980). Age_of_acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behaviour Research Methods and Instrumentation 12*, 395-427.

Goslin, J., & Frauenfelder, U. H. (2001). A comparison of theoretical and human syllabification. *Language and Speech*, *44*(4), 409–436.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*, 643-650.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition, 102*, 219-260.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*, 643-650.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, 1:174. doi: 10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavioral Research Methods, 44(1),* 287-304.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of Present-Day American English*. Providence, RI: Brown University Press.

Laganaro, M. (2008). Is there a syllable frequency effect in aphasia or in apraxia of speech or both?. *Aphasiology*, *22*(11), 1191-1200.

Laporte, E. (1993). Phonetic syllables in French: combinations, structure, and formal definitions. *Acta Linguistica Hungarica*, 41, 175-189.

Laudanna, A., Thornton, A.M., Brown, G., Burani, C., & Marconi, L. (1995). Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. In S. Bolasco, L. Lebart e A. Salem (a cura di), *III Giornate internazionali di Analisi Statistica dei Dati Testuali*. Volume I, pp. 103-109. Roma: Cisu

Maraschio, N. (1993). Grafia e ortografia: evoluzione e codificazione. In L. Serianni and P. Trifone (eds.), *Storia della lingua italiana* (vol. I, pp.139-227). Turin: Giulio Einaudi Editore.

Maturi, P. (2009). I suoni delle lingue, i suoni dell’italiano. Introduzione alla fonetica. Bologna: Il Mulino.

MacNeilage, P.F. (1982). Speech production mechanisms in aphasia. In S. Grillner, B., Lindblom, J. Lubker, & A. Persson (Eds.), *Speech Motor Control*, New York: Pergamon Press, pp 43-60.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28,* 661-677.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004) Lexique 2 : A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers, 36 (3),* 516-524.

New B., Pallier C., Ferrand L., & Matos R. (2001) Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Pschologique, 101,* 447-462.

Pérennou, G., & De Calmes, M. (1987). *BDLEX Base de données lexicales du français écrit et parlé*. Volume 1, Lexique général. : Travaux du Laboratoire CERFIA.

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources & Evaluation, 46*, 449-459. doi:10.1007/s10579-010-9130-z

Robson, J., Pring, T., Marshall, J., & Chiat, S. (2003). Phoneme frequency effects in jargon aphasia: A phonological investigation of nonword errors. Brain and Language, 85, 109-124.

Romani, C., Galluzzi, C., & Goslin, J. (submitted). A comparative study of phoneme frequency of use, age of acquisition and phonological complexity in Italian.

Romani, C., & Galluzzi, C. (2005). Effects of syllabic complexity in predicting accuracy of repetition and direction of errors in patients with articulatory and phonological difficulties. *Cognitive Neuropsychology, 22(7),* 817-850.

Romani, C., Granà, A., & Semenza, C. (1996). More errors on vowels than on consonants: An unusual case of conduction aphasia. *Brain and Language*, *55*(1), 144-146.

Romani, C., Galluzzi, C., Bureca, I., & Olson, A. (2011a). Effects of syllable structure in aphasic errors: Implications for a new model of speech production. Cognitive Psychology, 62, 151-192.

Romani, C., Galluzzi, C. & Olson, A. (2011b). Phonological lexical activation: A lexical component or an output buffer? Evidence from aphasic errors. Cortex, 47, 217-235.

Sebastián-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Una base de datos informatizada del español.* Barcelona: Servicio de Publicaciones de la Universitat de Barcelona.

Staiger, A. & Ziegler, W. (2008). Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech*. Aphasiology, 22,* 1201–1215.

Thorndike, E. L., & Lorge, I. (1944). *A teacher's word book of 30.000 words.* New York: Columbia University Press.

Wilshire, C. E., & Nespoulous, J. L. (2003). Syllables as units in speech production: Data from aphasia. *Brain and Language*, *84*(3), 424-447.

Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20(1),* 6-11.

	Field Name	Min	Max	Mean	SD
Frequency Fields					
Colfis Total Frequency	*fqTot*	0	119430	27.62	662.69
Log Colfis Total Frequency	*fqTotL*	0	11.69	1.19	1.39
General Orthographic Fields					
Number of letters	*NumLetters*	1	26.00	8.64	2.59
Consonant vowel structure of orthography	*OrthVCV*				
Orthgraphic uniqueness point	*OrthUniq*	0	18	6.52	3.97
Orthgraphic uniqueness point -1	*OrthUniqM1*	1	17	7.16	2.31
Number of Homographs	*NumHomographs*	0	25	0.71	1.55
General Phonological Fields					
Phonological representation of the word form	*Phones*				
Phonological representation with syllable boundary location (denoted by '.')	*PhonSyll*				
Number of phonemes	*NumPhones*	1	26	8.54	2.60
Consonant vowel structure of phonology	*PhonVCV*				
Number of syllables	*NumSylls*	1	11	3.66	1.11
Position of the stressed syllable	*StressedSyllable*	1	9	2.55	1.08
Phonological Uniqueness Point	*PhonUniq*	0	19	6.64	3.72
Phonological Uniqueness Point -1	*PhonUniqM1*	1	18	6.94	2.36
Number of Homophones	*NumHomophones*	0	41	0.76	1.89
Orthographic Neighbourhood and Levenshtein Distance Fields					
Orthographic neighbourhood size	*Orth_N*	0	28	2.31	3.02
Summed neighbourhood frequency	*Orth_N_MFreq*	0	11.16	1.35	1.45
Neighbourhood with greater frequency	*Orth_N_G*	0	24	1.32	2.18

Neighbourhood with lesser frequency	*Orth_N_L*	0	23	0.76	1.59
Summed frequency for neighbourhood of greater frequency	*Orth_N_G_MFreq*	0	11.35	1.50	1.83
Summed Frequency for neighbourhood of lesser frequency	*Orth_N_L_MFreq*	0	9.99	0.33	0.76
Relative log frequency between word and it's neighbourhood	*Orth_N_RelFreq*	0	30.22	0.51	0.92
Orthographic Levenshtein Distance 20	*OLD*	1	14.05	2.55	0.92
Summed frequency of words within OLD20	*OLDF*	0	6.69	1.70	0.69
Relative log frequency between word and those in the OLD20	*OLD_RelFreq*	0	10	0.70	0.79
Phonological Neighbourhood and Levenshtein Distance Fields					
Phonological neighbourhood size	*Phon_N*	0	30	2.29	2.93
Summed neighbourhood frequency	*Phon_N_MFreq*	0	10.36	1.37	1.46
Neighbourhood with greater frequency	*Phon_N_G*	0	26	1.32	2.14
Neighbourhood with lesser frequency	*Phon_N_L*	0	25	0.75	1.55
Summed frequency for neighbourhood of greater frequency	*Phon_N_G_MFreq*	0	11.46	1.51	1.83
Summed Frequency for neighbourhood of lesser frequency	*Phon_N_L_MFreq*	0	8.00	0.33	0.76
Relative log frequency between word and it's neighbourhood	*Phon_N_RelFreq*	0	28.25	0.52	0.92
Phonological Levenshtein Distance 20	*PLD*	1	14.55	2.60	0.94
Summed frequency of words within PLD20	*PLDF*	0.03	8.30	1.71	0.73

Table 1: Summary of *phonItalia main database fields and summarising statistics (where appropriate).*

Phone category	*Phone (IPA)*	*Phone (ascii)*	*TypeF*	*Proportion of TypeF*	*TokenF*	*Proportion of TokenF*	*LnTokenF*	*Proportion of LnTokenF*	*Example (orthographic)*	*Example (phonological)*
Vowels	a	a	130099	0.168	1998135	0.161	14.51	0.054	R**a**ta	/r**a**ta/
	i	i	102018	0.132	1494923	0.121	14.22	0.053	M**i**te	/m**i**te/
	o	o	84341	0.109	1417911	0.114	14.16	0.053	D**o**ve	/d**o**ve/
	e	e	81341	0.105	1555888	0.126	14.26	0.053	R**e**te	/r**e**te/
	u	u	17930	0.023	382939	0.031	12.86	0.048	M**u**to	/m**u**to/
	ɛ	E	14438	0.019	342453	0.028	12.74	0.048	M**e**ta	/m**E**ta/
	ɔ	O	9650	0.012	200376	0.016	12.21	0.046	**M**oto	/m**O**to/
Consonants	t	t	83848	0.108	1151501	0.093	13.96	0.052	**T**ana	/**t**ana/
	r	r	81414	0.105	1082468	0.087	13.9	0.052	**r**ete	/**r**ete/
	n	n	69115	0.089	1193267	0.096	13.99	0.052	**n**occa	/**n**Okka/
	s/z	s	55371	0.072	857307	0.069	13.67	0.051	**s**ano	/**s**ano/
	l	l	42387	0.055	898432	0.072	13.71	0.051	**l**ama	/**l**ama/
	k	k	39278	0.051	637446	0.051	13.37	0.05	**C**ane	/**k**ane/
	m	m	30659	0.04	446039	0.036	13.01	0.049	m**o**lla	/**m**Olla/
	p	p	27948	0.036	485715	0.039	13.09	0.049	**P**ane	/**p**ane/
	d	d	25764	0.033	594549	0.048	13.3	0.05	**D**anno	/**d**anno/
	v	v	19240	0.025	294196	0.024	12.6	0.047	**v**ano	/**v**ano/
	j	j	16525	0.021	249734	0.02	12.43	0.047	**i**eri	/**j**Eri/
	b	b	14666	0.019	165864	0.013	12.02	0.045	**B**anco	/**b**anko/
	f	f	14200	0.018	187581	0.015	12.14	0.045	**f**ame	/**f**ame/
	tʃ	c	13398	0.017	165300	0.013	12.02	0.045	**c**ena	/**c**ena/
	ts	z	12184	0.016	175804	0.014	12.08	0.045	**z**itto	/**z**itto/
	ʤ	g	10070	0.013	121624	0.01	11.71	0.044	**g**amba	/**g**amba/

g	G	9728	0.013	95160	0.008	11.47	0.043	**g**atto	/**G**atto/
w	w	5134	0.007	130437	0.011	11.78	0.044	**u**omo	/**w**Omo/
ʎ	L	4055	0.005	76278	0.006	11.24	0.042	**gl**i	/**L**i/
dz	Z	3944	0.005	25640	0.002	10.15	0.038	**z**ona	/**Z**Ona/
ʃ	S	3759	0.005	45706	0.004	10.73	0.04	**sc**endo	/**S**endo/
ɲ	N	3365	0.004	49064	0.004	10.81	0.04	o**gn**i	/o**NN**i/

Table 2: Summary of phone frequency of occurrences and the proportion of total frequency across the lexicon, ordered by type frequency

	Non-Geminate		*Geminate*		*Proportion of Geminates*	
Phone	*TypeF*	*TokenF*	*TypeF*	*TokenF*	*by TypeF*	*by TokenF*
r	76190	1030140	5224	52328	0.06	0.05
t	66926	896135	16922	255366	0.2	0.22
n	64579	1107587	4536	85680	0.07	0.07
s	43567	680079	11804	177228	0.21	0.21
k	31898	562654	7380	74792	0.19	0.12
l	31829	632544	10558	265888	0.25	0.3
m	27259	413993	3400	32046	0.11	0.07
d	24866	586463	898	8086	0.03	0.01
p	23834	436023	4114	49692	0.15	0.1
v	17826	278992	1414	15204	0.07	0.05
b	11658	112238	3008	53626	0.21	0.32
f	10760	152903	3440	34678	0.24	0.18
c	9454	133040	3944	32260	0.29	0.2
G	9214	92764	514	2396	0.05	0.03
g	5658	72456	4412	49168	0.44	0.4
z	2378	39240	9806	136564	0.8	0.78
S	561	6658	3198	39048	0.85	0.85
Z	492	4062	3452	21578	0.88	0.84
L	253	13440	3802	62838	0.94	0.82
N	13	62	3352	49002	1	1
All phones	*459215*	*7251473*	*105178*	*1497468*	*0.19*	*0.17*

Table 3: Summary of relative geminate and non-geminate frequency for consonants

Syllable Length	*TypeF*	*Proportion of TypeF*	*TokenF*	*Proportion of TokenF*
1	14251	0.032	602652	0.082
2	282385	0.642	4685270	0.634
3	126878	0.288	1877745	0.254
4	15307	0.035	219726	0.03
5	994	0.002	7222	0.001

Table 4: Summary of the frequency of use for syllables according to length

Number of Syllables	*Stressed Syllable* 1	2	3	4	5	6	7	8	9
1	1496 (*1134339*)								
2	13666 (*961482*)	1404 (*16641*)							
3	5377 (*119015*)	31090 (*557756*)	1189 (*10526*)						
4	186 (*1070*)	6688 (*69601*)	33554 (*287414*)	806 (*6447*)					
5	1 (*1*)	151 (*788*)	4144 (*21042*)	13968 (*97895*)	329 (*1299*)				
6	2 (*10*)	0 (*0*)	48 (*125*)	1443 (*5729*)	3148 (*18770*)	150 (*623*)			
7	0 (*0*)	0 (*0*)	0 (*0*)	10 (*13*)	249 (*807*)	678 (*2157*)	53 (*140*)		
8	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	1 (*1*)	31 (48)	105 (*321*)	8 (*24*)	
9	1 (*1*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	1 (*2*)	13 (*25*)	6 (*10*)
10	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	1 (*1*)	2 (*2*)
11	0 (*1*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	0 (*0*)	1 (*1*)

Table 5: Distribution of syllable stress by type frequency and (*token frequency*) according to the number of syllables in each word

	Syllable Onsets				*Syllable Codas*			
Length	*TypeF*	*Proportion of TypeF*	*TokenF*	*Proportion of TokenF*	*TypeF*	*Proportion of TypeF*	*TokenF*	*Proportion of TokenF*
0	37102	0.088	1144311	0.162	308073	0.70	5297255	0.717
1	353943	0.842	5492438	0.775	131367	0.299	2088909	0.283
2	28570	0.068	439182	0.062	372	0.001	6424	0.001
3	878	0.002	7724	0.001	5	0	37	0

Table 6: Summary of the frequency of use for syllable onsets and codas according to length

Phoneme	Freq in COLFIS	Ph-Apraxic Patients: N in corpus	Substitutions: Phoneme replaced N	Phoneme replaced %	Phoneme replacing N	Ph-Selection Patients: N in corpus	Substitutions: Phoneme replaced N	Phoneme replaced %	Phoneme replacing N
n	1,193,267	3817	94	**2.5**	61	3123	74	**2.4**	83
t	1,151,501	5214	87	**1.7**	400	4266	110	**2.6**	95
r	1,082,468	4840	242	**5.0**	123	3960	60	**1.5**	81
l	898,432	2849	129	**4.5**	242	2331	77	**3.3**	68
s	857,307	3015	114	**3.8**	68	2475	48	**1.9**	58
k	637,446	2475	127	**5.1**	199	2025	53	**2.6**	88
d	594,549	1320	199	**15.1**	84	1080	71	**6.6**	42
p	485,715	1936	90	**4.6**	245	1584	50	**3.2**	40
m	446,039	1936	64	**3.3**	40	1584	51	**3.2**	34
v	294,196	1045	188	**18.0**	66	855	52	**6.1**	33
j	249,734	1166	19	**1.6**	10	954	7	**0.7**	2
f	187,581	1342	135	**10.1**	123	1098	37	**3.4**	48
Z	175,804	770	39	**5.1**	47	630	19	**3.0**	19
b	165,864	891	122	**13.7**	129	729	23	**3.2**	27
c	165,300	814	67	**8.2**	98	666	18	**2.7**	28
w	130,437	462	9	**1.9**	2	378	3	**0.8**	0
g	121,624	418	74	**17.7**	10	342	15	**4.4**	6
G	95,160	726	179	**24.7**	19	594	32	**5.4**	32
Corr with freq				**-0.50**	**0.55**			**-0.22**	**0.87**
p				0.04	0.02			n.s.	<.001
confidence interval				-0.78 -- -0.04	0.11 -- 0.81			-0.63 -- 0.27	0.68 -- 0.95

Table 7: Substitution errors made by phonological-apraxic and phonological-selection aphasic patients.

	Total		*MonoSyll*		*Initial*		*Medial*		*Final*	
phones	*TypeF*	*TokenF*	*TypeF*	*TokenF*	*TypeF*	*TokenF*	*TypeF*	*TokenF*	*TypeF*	*TokenF*
to	12439	253020	4	40	128	1614	3535	27713	8772	223653
a	5288	205688	42	94559	2835	64799	1270	13988	1141	32342
di	4920	194778	22	130896	1615	27402	2557	24213	726	12267
ta	14202	179603	1	3	227	2510	6203	63687	7771	113403
la	5754	171998	25	65764	496	6629	3026	21177	2207	78428
ti	15476	160956	4	1612	282	3989	6958	68231	8232	87124
no	8274	141200	0	0	261	6744	883	5756	7130	128700
re	7911	136664	7	469	1098	15546	1272	6908	5534	113741
e	2791	125205	10	84690	2001	24844	311	3404	469	12267
te	10815	121077	11	785	472	7349	3047	26322	7285	86621
le	4656	114101	14	26163	330	3478	1340	10669	2972	73791
si	6101	104027	30	29368	341	11296	2465	28826	3265	34537
in	5077	103679	12	52861	4917	49813	143	805	5	200
ke	1322	99242	26	67238	27	55	340	1605	929	30344
ri	9284	98472	0	0	4108	38046	3062	32510	2114	27916
ra	6498	98240	2	2	441	7000	3870	35209	2185	56029
ne	5805	97371	12	4660	266	8339	1494	14313	4033	70059
na	6295	95352	3	21	226	4703	4010	33189	2056	57439
i	3068	90195	1	20	1179	19122	753	5911	1135	65142
ko	4784	84741	0	0	1009	34250	1895	23536	1880	26955
ma	3936	83359	11	17515	1173	21216	2193	21998	559	22630
so	2733	83181	7	690	568	31817	890	10276	1268	40398
E	329	81888	10	60538	178	18656	131	1674	10	1020
ka	7418	79722	3	36	1731	25475	3716	28754	1968	25457
ni	5810	74724	1	2	113	782	2225	24739	3471	49201
del	103	69922	10	32243	53	37489	40	190	0	0
kon	3339	69856	5	25760	2704	34952	628	9142	2	2
il	179	67947	9	66944	167	998	0	0	3	5

al	1182	67924	16	20230	1091	46053	71	1629	4	12
se	3801	66343	35	12860	785	13187	1190	13875	1791	26421
li	6459	65530	11	2118	554	9006	2716	24306	3178	30100
sa	3682	63354	9	889	659	16536	1845	18601	1169	27328
va	5111	63267	12	1796	412	5187	2592	25055	2095	31229
do	4848	62947	0	0	455	18422	2057	7975	2336	36550
de	3585	62221	19	2483	1520	31462	1510	15156	536	13120
lo	3446	60740	8	9810	143	6282	1128	10152	2167	34496
da	2517	60189	13	22900	190	9631	1640	16680	674	10978
per	999	58831	10	42143	576	14685	395	1794	18	209
mi	4163	57096	6	7140	816	19538	2170	21263	1171	9155
an	1408	52312	5	68	1279	50954	121	1283	3	7
un	82	52089	23	51498	55	434	3	156	1	1
ve	1895	49813	9	112	473	11359	983	27405	430	10937
ci	4146	48029	22	8489	402	4214	1858	21638	1864	13688
u	703	47172	0	0	646	46244	48	915	9	13
zjo	2725	46556	0	0	0	0	2576	42109	149	4447
po	1918	43780	2	74	517	15020	1187	10913	212	17773
mo	3929	43774	3	23	581	8382	1037	8243	2308	27126
me	1589	43594	13	2262	557	8580	718	10033	301	22719
ro	3404	42974	0	0	383	4849	1524	8849	1497	29276
o	1824	41521	16	8254	779	12534	588	4166	441	16567
men	2772	40823	0	0	69	3203	2698	37573	5	47
vi	2995	38245	15	1575	626	16326	1514	14975	840	5369
non	35	35710	4	35514	7	137	22	47	2	12
fi	2516	33203	0	0	599	15860	1819	16917	98	426
pa	2600	32732	0	0	1126	19137	1316	9766	158	3829
vo	2070	32044	0	0	267	7077	864	12410	939	12557
su	944	30144	17	4926	462	18976	449	5936	16	306
za	1648	28588	0	0	6	13	798	3590	844	24985

ce	1771	28221	13	1016	286	1898	874	9934	598	15373
tra	1790	27715	4	5083	802	6630	828	9200	156	6802
sta	271	26644	5	1644	229	24803	33	186	4	11
pre	1436	26135	1	2	1140	16563	279	4609	16	4961
bi	2459	24997	2	5	241	3194	2089	20125	127	1673
Li	504	24963	6	12501	0	0	37	114	461	12348
tro	864	23287	0	0	95	3269	520	2801	249	17217
tu	1881	22781	7	827	139	1910	1704	19749	31	295
pro	1579	22560	0	0	1231	20408	335	2021	13	131
nel	67	22042	6	12007	13	9830	44	153	4	52
pe	1541	21511	3	3	494	7282	903	12126	141	2100
gi	1774	20237	0	0	192	2721	1245	11257	337	6259
ku	1002	19916	0	0	233	7325	763	12582	6	9
fa	1030	19466	9	3605	436	12357	494	2723	91	781
par	742	19324	0	0	364	16308	373	3000	5	16
Ga	2063	19295	0	0	245	2181	1447	11879	371	5235
pju	57	17585	12	17053	11	58	27	438	7	36
go	841	17412	1	4	181	4873	358	5389	301	7146
be	1212	16539	0	0	248	1509	512	4961	452	10069
ca	1412	16030	1	32	21	104	1124	10665	266	5229
Go	1312	15943	0	0	156	2595	737	7881	419	5467
pi	1640	15748	0	0	257	1723	1191	10284	192	3741
du	812	15338	7	65	148	9241	647	6017	10	15
tan	906	15181	0	0	81	4562	808	10531	17	88
tut	79	14864	1	2	60	12952	18	1910	0	0
pri	449	14632	0	0	213	12234	209	1841	27	557
kwes	48	14602	0	0	25	14439	23	163	0	0
pO	216	14493	10	2974	80	10223	108	1234	18	62
dal	50	14458	5	6897	39	7547	2	3	4	11
im	1584	14065	0	0	1567	14040	17	25	0	0

ki	1023	14013	4	3400	94	1488	431	2820	494	6305
kom	1070	13381	7	263	877	12025	184	1083	2	10
tri	1199	12673	1	1	211	1434	830	5022	157	6216
lu	887	12350	2	4	256	6991	616	5315	13	40
kwel	15	12340	5	2606	10	9734	0	0	0	0
ge	1326	12003	0	0	294	3722	863	5252	169	3029
sul	148	11933	4	4556	20	5394	124	1983	0	0
ar	1094	11635	0	0	1005	11413	85	216	4	6
nu	638	11534	0	0	133	2491	497	8750	8	293
tre	307	11521	3	2811	86	608	123	643	95	7459
sjo	684	11400	0	0	0	0	613	11145	71	255
kwa	342	11386	7	243	193	9569	128	499	14	1075

Appendix A: 100 most frequent syllables (by token) with frequency of occurrence data across the entire lexicon and relative to word position (monosyllables, word initial, medial, and final position).