Behav Res (2014) 46:872–886
DOI 10.3758/s13428-013-0400-8

PhonItalia: a phonological lexicon for Italian

Jeremy Goslin · Claudia Galluzzi · Cristina Romani

Published online: 4 October 2013

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 includes 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, and character and phoneme bigrams. These databases are freely available from *phonitalia.org*. This article describes the methods, content, and summarizing 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.

Keywords Phonological lexicon · Lexical statistics · Aphasic errors

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 behavioral 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 modeling 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 (Kučera & Francis, 1967; Thorndike & Lorge, 1944) 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 behavioral 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 Calmès & Pérennou, 1998; Lexique, New, Pallier, Brysbaert, & Ferrand, 2004; New, Pallier, Ferrand, & Matos, 2001; CELEX, Baayen, Piepenbrock, & van Rijn, 1993) and recent behavioral 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)

J. Goslin (✉)
School of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK
e-mail: jeremy.goslin@plymouth.ac.uk

C. Galluzzi
Fondazione Santa Lucia, i.r.c.s.s., Rome, Italy

C. Romani
School of Life and Health Sciences, Aston University, Birmingham, UK

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 (Bertinetto et al., 2005; Laudanna, Thornton, Brown, Burani, & Marconi, 1995) 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 2,719 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 provide 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 and phoneme that could make word-level phonological transcription largely redundant. However, while Italian can be classified as being toward 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" and "e." In contrast, the affricate /ʤ/ is realized by the letter "g" if followed by the vowels "i" and "e," but by the bigram "gi" if followed by the vowels /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 gemination 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 article, we 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 article 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 aphasics' 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, and Goslin (2013).

Methodology

The basis for this lexicon was Colfis (Bertinetto et al., 2005; Laudanna et al., 1995), 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

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

[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).

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 nonwords (such as "-se-") and minor corrections to 2,294 word forms (e.g., 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 nonwords (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 labeled 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 of 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 that 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 that 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, 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 native Italian 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 color their judgments. 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 2,917 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 was 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. Since the uniqueness point lists a value of zero if the representation never becomes unique, an additional field is also included that lists the uniqueness point minus one (*OrthUniqM1*). For nonunique words, this field will have a value of the length of the word and, thus, avoids the potential skewing in summarizing 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 neighbor. 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/phonological representations 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 is 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).

For 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.

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

Derived sublexical 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 nonpositional 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 * 2,408). 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 multiconsonantal syllable onsets are provided (e.g., the frequency of onsets like, /p/, /pr/, /pl/, or /str/). Syllable frequencies are 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 summarized in the following sections. As with the main lexicon, all these additional databases are available from the lexicon Web site in Excel and tab-delimited text format. The source code 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 Web site *phonItalia.org*.

Phone statistics

This database provides the frequency of occurrence for all 29 Italian phones used within this lexicon. Overall phonemic frequencies of use are summarized 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/.

Table 1 Summary of *phonItalia* main database fields and summarizing statistics (where appropriate)

	Field Name	Min	Max	Mean	SD
Frequency Fields					
Colfis total frequency	*fqTot*	0	1,19430	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 Neighborhood and Levenshtein Distance Fields					
Orthographic neighborhood size	*Orth_N*	0	28	2.31	3.02
Summed neighborhood frequency	*Orth_N_MFreq*	0	11.16	1.35	1.45
Neighborhood with greater frequency	*Orth_N_G*	0	24	1.32	2.18
Neighborhood with lesser frequency	*Orth_N_L*	0	23	0.76	1.59
Summed frequency for neighborhood of greater frequency	*Orth_N_G_MFreq*	0	11.35	1.50	1.83
Summed frequency for neighborhood of lesser frequency	*Orth_N_L_MFreq*	0	9.99	0.33	0.76
Relative log frequency between word and its neighborhood	*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 Neighborhood and Levenshtein Distance Fields					
Phonological neighborhood size	*Phon_N*	0	30	2.29	2.93
Summed neighborhood frequency	*Phon_N_MFreq*	0	10.36	1.37	1.46
Neighborhood with greater frequency	*Phon_N_G*	0	26	1.32	2.14
Neighborhood with lesser frequency	*Phon_N_L*	0	25	0.75	1.55
Summed frequency for neighborhood of greater frequency	*Phon_N_G_MFreq*	0	11.46	1.51	1.83
Summed frequency for neighborhood of lesser frequency	*Phon_N_L_MFreq*	0	8.00	0.33	0.76
Relative log frequency between word and its neighborhood	*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

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 nucleus in /a.E.r**o**.pla.n**o**/.

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

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	130,099	.168	1,998,135	.161	14.51	.054	R**a**ta	/r**a**ta/
	i	i	102,018	.132	1,494,923	.121	14.22	.053	M**i**te	/m**i**te/
	o	o	84,341	.109	1,417,911	.114	14.16	.053	D**o**ve	/d**o**ve/
	e	e	81,341	.105	1,555,888	.126	14.26	.053	R**e**te	/r**e**te/
	u	u	17,930	.023	382,939	.031	12.86	.048	M**u**to	/m**u**to/
	ɛ	E	14,438	.019	342,453	.028	12.74	.048	M**e**ta	/m**E**ta/
	ɔ	O	9,650	.012	200,376	.016	12.21	.046	M**o**to	/m**O**to/
Consonants	t	t	83,848	.108	1,151,501	.093	13.96	.052	**T**ana	/**t**ana/
	r	r	81,414	.105	1,082,468	.087	13.9	.052	**r**ete	/**r**ete/
	n	n	69,115	.089	1,193,267	.096	13.99	.052	**n**occa	/**n**Okka/
	s/z	s	55,371	.072	857,307	.069	13.67	.051	**s**ano	/**s**ano/
	l	l	42,387	.055	898,432	.072	13.71	.051	**l**ama	/**l**ama/
	k	k	39,278	.051	637,446	.051	13.37	.05	**C**ane	/**k**ane/
	m	m	30,659	.04	446,039	.036	13.01	.049	m**o**lla	/**m**Olla/
	p	p	27,948	.036	485,715	.039	13.09	.049	**P**ane	/**p**ane/
	d	d	25,764	.033	594,549	.048	13.3	.05	**D**anno	/**d**anno/
	v	v	19,240	.025	294,196	.024	12.6	.047	**v**ano	/**v**ano/
	j	j	16,525	.021	249,734	.02	12.43	.047	**i**eri	/**j**Eri/
	b	b	14,666	.019	165,864	.013	12.02	.045	**B**anco	/**b**anko/
	f	f	14,200	.018	187,581	.015	12.14	.045	**f**ame	/**f**ame/
	ʧ	c	13,398	.017	165,300	.013	12.02	.045	**c**ena	/**c**ena/
	ts	z	12,184	.016	175,804	.014	12.08	.045	**z**itto	/**z**itto/
	ʤ	g	10,070	.013	121,624	.01	11.71	.044	**g**amba	/**g**amba/
	g	G	9,728	.013	95,160	.008	11.47	.043	**g**atto	/**G**atto/
	w	w	5,134	.007	130,437	.011	11.78	.044	**u**omo	/**w**Omo/
	ʎ	L	4,055	.005	76,278	.006	11.24	.042	**gl**i	/**L**i/
	dz	Z	3,944	.005	25,640	.002	10.15	.038	**z**ona	/**Z**Ona/
	ʃ	S	3,759	.005	45,706	.004	10.73	.04	**sc**endo	/**S**endo/
	ɲ	N	3,365	.004	49,064	.004	10.81	.04	o**gn**i	/o**NN**i/

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 nongeminate (e.g., /n/ in /pu**n**.to/ "punto").

Syllable statistics

This database contains calculations of the frequency of use for the 3,626 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 do orthographic syllables, with only 2,719 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

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

Table 3 Summary of relative geminate and nongeminate frequencies for consonants

Phone	Nongeminate		Geminate		Proportion of Geminates	
	TypeF	TokenF	TypeF	TokenF	by TypeF	by TokenF
r	76,190	1,030,140	5,224	52,328	.06	.05
t	66,926	896,135	16,922	255,366	.2	.22
n	64,579	1,107,587	4,536	85,680	.07	.07
s	43,567	680,079	11,804	177,228	.21	.21
k	31,898	562,654	7,380	74,792	.19	.12
l	31,829	632,544	10,558	265,888	.25	.3
m	27,259	413,993	3,400	32,046	.11	.07
d	24,866	586,463	898	8,086	.03	.01
p	23,834	436,023	4,114	49,692	.15	.1
v	17,826	278,992	1,414	15,204	.07	.05
b	11,658	112,238	3,008	53,626	.21	.32
f	10,760	152,903	3,440	34,678	.24	.18
c	9,454	133,040	3,944	32,260	.29	.2
G	9,214	92,764	514	2,396	.05	.03
g	5,658	72,456	4,412	49,168	.44	.4
z	2,378	39,240	9,806	136,564	.8	.78
S	561	6,658	3,198	39,048	.85	.85
Z	492	4,062	3,452	21,578	.88	.84
L	253	13,440	3,802	62,838	.94	.82
N	13	62	3,352	49,002	1	1
All phones	*459,215*	*7,251,473*	*105,178*	*1,497,468*	*.19*	*.17*

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:

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/.

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

Syllable Length	TypeF	Proportion of TypeF	TokenF	Proportion of TokenF
1	14,251	.032	602,652	.082
2	282,385	.642	4,685,270	.634
3	126,878	.288	1,877,745	.254
4	15,307	.035	219,726	.03
5	994	.002	7,222	.001

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 Table 8, 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 sequences 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 includes the statistics only 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/.

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

Number of Syllables	Stressed Syllable								
	1	2	3	4	5	6	7	8	9
1	1,496 (*1,134,339*)								
2	13,666 (*961,482*)	1,404 (*16,641*)							
3	5,377 (*119,015*)	31,090 (*557,756*)	1,189 (*10,526*)						
4	186 (*1,070*)	6,688 (*69,601*)	33,554 (*287,414*)	806 (*6,447*)					
5	1 (*1*)	151 (*788*)	4,144 (*21,042*)	13,968 (*97,895*)	329 (*1,299*)				
6	2 (*10*)	0 (*0*)	48 (*125*)	1,443 (*5,729*)	3,148 (*18,770*)	150 (*623*)			
7	0 (*0*)	0 (*0*)	0 (*0*)	10 (*13*)	249 (*807*)	678 (*2,157*)	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*)

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, summarized 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 for both syllable onsets and codas that are medial to the word—for example, the onset /d/ or the coda /n/ in /mo**n**.**d**o/.
Word Final provides statistics 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/.

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

Length	Syllable Onsets				Syllable Codas			
	TypeF	Proportion of TypeF	TokenF	Proportion of TokenF	TypeF	Proportion of TypeF	TokenF	Proportion of TokenF
0	37,102	.088	1,144,311	.162	308,073	.70	5,297,255	.717
1	353,943	.842	5,492,438	.775	131,367	.299	2,088,909	.283
2	28,570	.068	439,182	.062	372	.001	6,424	.001
3	878	.002	7,724	.001	5	0	37	0

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 /s/ 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 phones 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 nor 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 phones 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 precedes 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 for diagnosing 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 ($n = \sim 6$) of Broca, Wernicke's, and conduction aphasics. However, a larger study by MacNeilage (1982) contrasted 20 English-speaking nonfluent aphasics (with possible apraxic difficulties) with 10 fluent aphasics. He found that target error rates were greater in low- than in high-frequency phonemes (frequency correlated with percentage of errors), but only in the nonfluent 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, Marshall, 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 postlexical locus, but not in a patient with lexical phonological impairment.

An examination of the limited evidence from these studies suggests 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 subgroups 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, 2005; Romani, Galluzzi, Bureca, & Olson, 2011; Romani, Galluzzi, & Olson, 2011). 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, since 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, Galluzzi, Bureca, et al., 2011, 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, $r = -.50$, $p < .05$, but no significant correlation in the ph-selection patients, $r = -.22$, $p = .36$. 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, $r = .55$, $p < .05$, and the ph-selection, $r = .87$, $p < .001$, patient groups. 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 rate 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 toward 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; Staiger & Ziegler, 2008), but not in patients with more central phonological difficulties (Wilshire & Nespoulous, 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 and characterize apraxic patients (Romani & Galluzzi, 2005; Romani, Galluzzi, Bureca, et al., 2011; Romani, Granà, & Semenza, 1996). Both analyses of frequency and complexity highlight important differences between types of patients that are not well recognized 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, Guariglia & Romani, 2013). It is possible, however, that, although strongly related,

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

Phoneme	Freq in COLFIS	Ph-Apraxic Patients				Ph-Selection Patients			
			Substitutions				Substitutions		
		N in Corpus	Phoneme Replaced		Phoneme Replacing	*N* in Corpus	Phoneme Replaced		Phoneme Replacing
			N	%	*N*		*N*	%	*N*
n	1,193,267	3,817	94	2.5	61	3,123	74	2.4	83
t	1,151,501	5,214	87	1.7	400	4,266	110	2.6	95
r	1,082,468	4,840	242	5.0	123	3,960	60	1.5	81
l	898,432	2,849	129	4.5	242	2,331	77	3.3	68
s	857,307	3,015	114	3.8	68	2,475	48	1.9	58
k	637,446	2,475	127	5.1	199	2,025	53	2.6	88
d	594,549	1,320	199	15.1	84	1,080	71	6.6	42
p	485,715	1,936	90	4.6	245	1,584	50	3.2	40
m	446,039	1,936	64	3.3	40	1,584	51	3.2	34
v	294,196	1,045	188	18.0	66	855	52	6.1	33
j	249,734	1,166	19	1.6	10	954	7	0.7	2
f	187,581	1,342	135	10.1	123	1,098	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				−.50	.55			−.22	.87
p				.04	.02			n.s.	<.001
Confidence interval				−0.78—0.04	0.11–0.81			−0.63–0.27	0.68–0.95

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 et al., (2013), 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 modalities. 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 Web site *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 and in natural language processing and 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 as to the locus of an individual's language impairment.

Appendix

Table 8 100 most frequent syllables (by token) with frequency of occurrence data across the entire lexicon and relative to word position (monosyllables and word-initial, -medial, and -final positions

Phones	Total		MonoSyll		Initial		Medial		Final	
	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF
to	12,439	253,020	4	40	128	1,614	3,535	27,713	8,772	223,653
a	5,288	205,688	42	94,559	2,835	64,799	1,270	13,988	1,141	32,342
di	4,920	194,778	22	130,896	1,615	27,402	2,557	24,213	726	12,267
ta	14,202	179,603	1	3	227	2,510	6,203	63,687	7,771	113,403
la	5,754	171,998	25	65,764	496	6,629	3,026	21,177	2,207	78,428
ti	15,476	160,956	4	1,612	282	3,989	6,958	68,231	8,232	87,124
no	8,274	141,200	0	0	261	6,744	883	5,756	7,130	128,700
re	7,911	136,664	7	469	1,098	15,546	1,272	6,908	5,534	113,741
e	2,791	125,205	10	84,690	2,001	24,844	311	3,404	469	12,267
te	10,815	121,077	11	785	472	7,349	3,047	26,322	7,285	86,621
le	4,656	114,101	14	26,163	330	3,478	1,340	10,669	2,972	73,791
si	6,101	104,027	30	29,368	341	11,296	2,465	28,826	3,265	34,537
in	5,077	103,679	12	52,861	4,917	49,813	143	805	5	200
ke	1,322	99,242	26	67,238	27	55	340	1,605	929	30,344
ri	9,284	98,472	0	0	4,108	38,046	3,062	32,510	2,114	27,916
ra	6,498	98,240	2	2	441	7,000	3,870	35,209	2,185	56,029
ne	5,805	97,371	12	4,660	266	8,339	1,494	14,313	4,033	70,059
na	6,295	95,352	3	21	226	4,703	4,010	33,189	2,056	57,439
i	3,068	90,195	1	20	1,179	19,122	753	5,911	1,135	65,142
ko	4,784	84,741	0	0	1,009	34,250	1,895	23,536	1,880	26,955
ma	3,936	83,359	11	17,515	1,173	21,216	2,193	21,998	559	22,630
so	2,733	83,181	7	690	568	31,817	890	10,276	1,268	40,398
E	329	81,888	10	60,538	178	18,656	131	1,674	10	1,020
ka	7,418	79,722	3	36	1,731	25,475	3,716	28,754	1,968	25,457
ni	5,810	74,724	1	2	113	782	2,225	24,739	3,471	49,201
del	103	69,922	10	32,243	53	37,489	40	190	0	0
kon	3,339	69,856	5	25,760	2,704	34,952	628	9,142	2	2
il	179	67,947	9	66,944	167	998	0	0	3	5
al	1,182	67,924	16	20,230	1,091	46,053	71	1,629	4	12
se	3,801	66,343	35	12,860	785	13,187	1,190	13,875	1,791	26,421
li	6,459	65,530	11	2,118	554	9,006	2,716	24,306	3,178	30,100
sa	3,682	63,354	9	889	659	16,536	1,845	18,601	1,169	27,328
va	5,111	63,267	12	1,796	412	5,187	2,592	25,055	2,095	31,229
do	4,848	62,947	0	0	455	18,422	2,057	7,975	2,336	36,550
de	3,585	62,221	19	2,483	1,520	31,462	1,510	15,156	536	13,120
lo	3,446	60,740	8	9,810	143	6,282	1,128	10,152	2,167	34,496
da	2,517	60,189	13	22,900	190	9,631	1,640	16,680	674	10,978
per	999	58,831	10	42,143	576	14,685	395	1,794	18	209
mi	4,163	57,096	6	7,140	816	19,538	2,170	21,263	1,171	9,155
an	1,408	52,312	5	68	1,279	50,954	121	1,283	3	7
un	82	52,089	23	51,498	55	434	3	156	1	1
ve	1,895	49,813	9	112	473	11,359	983	27,405	430	10,937

Table 8 (continued)

Phones	Total		MonoSyll		Initial		Medial		Final	
	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF
ci	4,146	48,029	22	8,489	402	4,214	1,858	21,638	1,864	13,688
u	703	47,172	0	0	646	46,244	48	915	9	13
zjo	2,725	46,556	0	0	0	0	2,576	42,109	149	4,447
po	1,918	43,780	2	74	517	15,020	1,187	10,913	212	17,773
mo	3,929	43,774	3	23	581	8,382	1,037	8,243	2,308	27,126
me	1,589	43,594	13	2,262	557	8,580	718	10,033	301	22,719
ro	3,404	42,974	0	0	383	4,849	1,524	8,849	1,497	29,276
o	1,824	41,521	16	8,254	779	12,534	588	4,166	441	16,567
men	2,772	40,823	0	0	69	3,203	2,698	37,573	5	47
vi	2,995	38,245	15	1,575	626	16,326	1,514	14,975	840	5,369
non	35	35,710	4	35,514	7	137	22	47	2	12
fi	2,516	33,203	0	0	599	15,860	1,819	16,917	98	426
pa	2,600	32,732	0	0	1,126	19,137	1,316	9,766	158	3,829
vo	2,070	32,044	0	0	267	7,077	864	12,410	939	12,557
su	944	30,144	17	4,926	462	18,976	449	5,936	16	306
za	1,648	28,588	0	0	6	13	798	3,590	844	24,985
ce	1,771	28,221	13	1,016	286	1,898	874	9,934	598	15,373
tra	1,790	27,715	4	5,083	802	6,630	828	9,200	156	6,802
sta	271	26,644	5	1,644	229	24,803	33	186	4	11
pre	1,436	26,135	1	2	1,140	16,563	279	4,609	16	4,961
bi	2,459	24,997	2	5	241	3,194	2,089	20,125	127	1,673
Li	504	24,963	6	12,501	0	0	37	114	461	12,348
tro	864	23,287	0	0	95	3,269	520	2,801	249	17,217
tu	1,881	22,781	7	827	139	1,910	1,704	19,749	31	295
pro	1,579	22,560	0	0	1,231	20,408	335	2,021	13	131
nel	67	22,042	6	12,007	13	9,830	44	153	4	52
pe	1,541	21,511	3	3	494	7,282	903	12,126	141	2,100
gi	1,774	20,237	0	0	192	2,721	1,245	11,257	337	6,259
ku	1,002	19,916	0	0	233	7,325	763	12,582	6	9
fa	1,030	19,466	9	3,605	436	12,357	494	2,723	91	781
par	742	19,324	0	0	364	16,308	373	3,000	5	16
Ga	2,063	19,295	0	0	245	2,181	1,447	11,879	371	5,235
pju	57	17,585	12	17,053	11	58	27	438	7	36
go	841	17,412	1	4	181	4,873	358	5,389	301	7,146
be	1,212	16,539	0	0	248	1,509	512	4,961	452	10,069
ca	1,412	16,030	1	32	21	104	1,124	10,665	266	5,229
Go	1,312	15,943	0	0	156	2,595	737	7,881	419	5,467
pi	1,640	15,748	0	0	257	1,723	1,191	10,284	192	3,741
du	812	15,338	7	65	148	9,241	647	6,017	10	15
tan	906	15,181	0	0	81	4,562	808	10,531	17	88
tut	79	14,864	1	2	60	12,952	18	1,910	0	0
pri	449	14,632	0	0	213	12,234	209	1,841	27	557
kwes	48	14,602	0	0	25	14,439	23	163	0	0
pO	216	14,493	10	2,974	80	10,223	108	1,234	18	62
dal	50	14,458	5	6,897	39	7,547	2	3	4	11
im	1,584	14,065	0	0	1,567	14,040	17	25	0	0
ki	1,023	14,013	4	3,400	94	1,488	431	2,820	494	6,305

Table 8 (continued)

Phones	Total		MonoSyll		Initial		Medial		Final	
	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF	TypeF	TokenF
kom	1,070	13,381	7	263	877	12,025	184	1,083	2	10
tri	1,199	12,673	1	1	211	1,434	830	5,022	157	6,216
lu	887	12,350	2	4	256	6,991	616	5,315	13	40
kwel	15	12,340	5	2,606	10	9,734	0	0	0	0
ge	1,326	12,003	0	0	294	3,722	863	5,252	169	3,029
sul	148	11,933	4	4,556	20	5,394	124	1,983	0	0
ar	1,094	11,635	0	0	1,005	11,413	85	216	4	6
nu	638	11,534	0	0	133	2,491	497	8,750	8	293
tre	307	11,521	3	2,811	86	608	123	643	95	7,459
sjo	684	11,400	0	0	0	0	613	11,145	71	255
kwa	342	11,386	7	243	193	9,569	128	499	14	1,075

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., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459.

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, and Computers, 34*(3), 424–434.

Bertinetto 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 & S. Blumstein (Eds.), *Psycholinguistics and aphasia* (pp. 123–236). Baltimore: Johns Hopkins University Press.

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

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.

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

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

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.

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.

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.

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

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... 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.

Galluzzi, C., Bureca, I., Guariglia, C., & Romani, C. (2013). Phonological simplifications and the differential diagnosis of apraxia of speech. Manuscript submitted for publication.

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.

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

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

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.

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

Maraschio, N. (1993). Grafia e ortografia: Evoluzione e codificazione. In L. Serianni & 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.

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

Romani, C., Galluzzi, C., & Goslin, J. (2013). A comparative study of phoneme frequency of use, age of acquisition and phonological complexity in Italian. Manuscript submitted for publication.

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

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.

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.

Stella, V., & Job, R. (2001). Le sillabe PD/DPSS. Una base di dati sulla frequenza sillabicadell'italiano scritto. *Giornale Italiano di Psicologia, 28*, 633–639.

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.