Q2Stress: A database for multiple cues to stress assignment in Italian

Giacomo Spinelli[1,^], Simone Sulpizio[2,3,^], & Cristina Burani[4,5]

[1]Department of Psychology, University of Western Ontario

[2]Faculty of Psychology, Vita-Salute San Raffaele University, Milano

[3]Department of Psychology and Cognitive Science, University of Trento

[4]Institute of Cognitive Science and Technology, CNR, Roma

[5]Department of Life Sciences, University of Trieste

^G.S. and S.S. contributed equally to this work as joint first authors

Running head: A database for stress assignment in Italian

Word count: 7,845

Corresponding author:
Giacomo Spinelli,
Department of Psychology, Social Science Centre, University of Western Ontario, London, ON N6A 5C2, Canada
E-mail: gspinel@uwo.ca

Abstract

In languages where the position of lexical stress within a word is not predictable from print, readers rely on distributional information extracted from the lexicon in order to assign stress. Lexical databases are thus especially important for researchers willing to address stress assignment in those languages. Here we present *Q2Stress*, a new database aimed to fill the lack of such a resource for Italian. *Q2Stress* includes multiple cues readers may use in assigning stress, such as type and token frequency of stress patterns as well as their distribution with respect to number of syllables, grammatical category, word beginnings, word endings, and consonant-vowel structures. Furthermore, for the first time, data for both adults and children are available. *Q2Stress* may help researchers to answer empirical as well as theoretical questions about stress assignment and stress related issues, and more in general, to explore the orthography-to-phonology relation in reading. *Q2Stress* is designed as a user-friendly resource, as it comes with scripts allowing researchers to explore and select their own stimuli according to several criteria as well as summary tables for overall data analysis.

Keywords: stress assignment, database, Italian, reading, word recognition

Mapping orthography onto phonology is the basic skill underlying literacy. A full understanding of how this is done is the main challenge for researchers interested in the interface between orthography and phonology. Most of empirical investigation has focused on segmental material and addressed the relationship between graphemes and phonemes; in recent years, however, researchers have started to properly consider information specified at the suprasegmental level. The assignment of stress, in particular, has proved to be especially challenging for models of polysyllabic word recognition and reading (e.g., Pagliuca & Monaghan, 2010; Perry, Ziegler, & Zorzi, 2010; Rastle & Coltheart, 2000).

In languages where stress position is not fixed, such as English, Spanish or Italian, stress may not be predictable from print in a straightforward manner, i.e., by means of rules alone. Rather, readers may use multiple sources of information in order to assign stress, including lexical look-up, distributional information, and grammatical category. For instance, consider the English word *JUStice* (note 1). Where does information concerning its stress pattern come from? In the first place, readers may recall that *JUStice* is stressed on the first syllable – rather than the second syllable, as in *poLICE* – by retrieving its entry in the lexicon (Perry et al., 2010; Rastle & Coltheart, 2000). Second, they may rely on distributional information. In English, first syllable stress is the dominant pattern and may act as a default (Brown, Lupker, & Colombo, 1994; Cutler & Carter, 1987; Monsell, Doyle, & Haggard, 1989; Rastle & Coltheart, 2000). Besides, first syllable stress is also the typical pattern of English nouns – verbs, on the contrary, show a bias towards second syllable stress, as in *adVICE* (Arciuli & Cupples, 2006; Kelly & Bock, 1988; Sereno, 1986). Finally, most English words ending in *-ice* bear stress on the first syllable (Arciuli & Cupples, 2006).

In the last thirty years, several studies have shed light on the role of different cues to stress assignment in free-stress languages such as English or Italian and how they interact within the reading system. However, we would not have made much progress without the aid of corpus analysis. With the growing interest in statistical learning approaches, lexical databases have turned out to be a fundamental support for research on stress (as well as many other linguistic aspects): As a source of

distributional information, databases have played a crucial role in driving hypotheses about stress assignment and inspiring behavioral and modeling studies. The availability of a lexical database is a prerequisite for those who are interested in research on stress. Quite disappointingly, however, this support is not available for some languages. For English, researchers can count on several resources – most notably CELEX (Baayen, Piepenbrock, & van Rijn, 1993) and the analyses made by Howard and Smith (2002) and Arciuli and Cupples (2006; 2007). Despite the remarkable number of studies concerned with lexical stress (see Sulpizio, Burani, & Colombo, 2015, for a review), no useful database for the investigation of stress assignment in Italian is currently available to the scientific community. Information concerning the distribution of stress patterns is also absent from (Italian) child corpora. With the present paper, we aim to fill this gap.

We begin by briefly reviewing the findings from the studies that addressed stress assignment in Italian and the resources available so far. Then, we present *Q2stress*, a lexical resource that provides information for multiple variables related to stress and constitutes a powerful tool for researchers interested in investigating stress issues in word recognition and reading. The following section reports the main results concerning the distribution of stress patterns in both adult and child-directed Italian corpora. Finally, we discuss some of the implications these data have for future research.

Stress assignment in Italian

Italian is especially appealing for research on stress. Like Spanish and unlike English, it is highly transparent at the segmental level, i.e., there is an almost one-to-one correspondence between graphemes and phonemes. Like English and unlike Spanish, however, Italian is highly opaque at the suprasegmental level, as in most cases stress patterns cannot be derived from orthographic rules or stress diacritics (note 2). Thus, on the one hand, stress in Italian can be studied rather independently from segmental phonology. On the other hand, the fact that in most cases stress is unpredictable allows researchers to inspect carefully what happens when no straightforward way to assign stress is available. Inquiring cues to stress is, in fact, much harder for languages where orthographic rules

apply, as rules (e.g., diacritics that mark the stressed syllable) may cover the vast majority of word forms (for Greek see, e.g., Protopapas, Gerakaki, & Alexandri, 2006; 2007). In sum, Italian presents a valuable opportunity to inquire how stress processing occurs in the reading system.

Research addressing reading aloud and word recognition in Italian has highlighted several factors involved in stress assignment. Similar to English, word knowledge is a primary source of information for assigning stress to known words (Colombo, 1992). Syllabic structure also plays some role, as words with a closed penultimate syllable (i.e., ending with a consonant) bear penultimate stress, although some exceptions apply (e.g., *MANdorla* 'almond', *FINferli* 'chanterelles', *POlizza* 'policy'). This behavior will be referred to as Italian's phonological rule.

Most studies investigating stress assignment in reading focused on distributional information, which can be represented at different levels of specificity. At the most general level, there is a dominant stress pattern in Italian: most words bear penultimate stress (i.e., stress on the penultimate syllable). In reading aloud, beginning readers show a bias towards the dominant penultimate stress (Colombo, Deguchi, & Boureux, 2014; Sulpizio & Colombo, 2013); with the increase of reading skills, however, the bias disappears, since skilled readers rely on a more specific source of information, namely stress neighborhood (Burani & Arduino, 2004; Burani, Paizi, & Sulpizio, 2014; Sulpizio, Arduino, Paizi, & Burani, 2013). Stress neighborhood is the proportion of words sharing the same orthographic word ending and stress pattern on the total number of words with that ending. For instance, the stress neighborhood for the word ending *-ola* is mainly associated with antepenultimate stress, since most words ending in *-ola* bear antepenultimate stress. Because of its mainly antepenultimate stress neighborhood and despite its non-dominant stress, an antepenultimate stress word ending in *-ola*, such as *PENtola* 'pot', is read faster and more accurately than a penultimate stress word with the same ending, such as *piSTOla* 'gun' (Burani & Arduino, 2004; Burani et al., 2014; Colombo, 1992). Stress neighborhood affects also nonword reading: A nonword ending in *-ola* is more likely to be assigned antepenultimate than penultimate stress (Sulpizio et al., 2013).

Taken together, findings from reading aloud indicate that, with a limited lexicon and weak orthography-to-phonology connections, younger readers assign stress mostly relying on general distributional information (i.e., stress dominance). As reading skills increase, general information is replaced by specific distributional information (i.e., stress neighborhood), which becomes the main source of information for stress assignment. In contrast, research on visual word recognition offers a very different picture. Studies conducted on adult readers suggest that stress neighborhood may play a weak role in lexical decision tasks (Burani & Arduino, 2004; Colombo & Sulpizio, 2015), whereas an advantage for (dominant) penultimate stress words was reported (Colombo, 1992; Colombo & Sulpizio, 2015). In sum, stress distributional information is specified at different levels, and both different stages of development and different tasks may recruit such information at different levels.

Cues to stress

While the literature focused on stress dominance and stress neighborhood, little is known concerning other factors that may be involved in stress assignment, such as grammatical category and word beginnings. Some studies conducted in languages other than Italian suggest that both grammatical category and word beginnings may have a role in driving stress assignment. Using both reading aloud and lexical decision tasks, Arciuli and Cupples (2006) and Jouravlev and Lupker (2014) found that grammatical category interacts with stress neighborhood in English and Russian, respectively (for English see also Kelly, 1988; Kelly & Bock, 1988; Smith, Baker & Groat, 1982). Arciuli and collaborators also found that not only word endings but also word beginnings affect English stress assignment in both adults (Arciuli & Cupples, 2007) and children (Arciuli, Monaghan, & Seva, 2010). According to a recent corpus analysis (Monaghan, Arciuli, & Seva, 2016), word beginnings, though less predictive than word endings, are good predictors of stress patterns in several languages, including Italian. Importantly, the role of both beginnings and endings in stress assignment seems to be independent from the morphological properties of those units (Arciuli & Cupples, 2006, 2007). That said, it is known that affixes either repel or attract stress (e.g., Jarmulowicz, Taran, & Hay, 2008; Ktori, Tree, Mousikou, Coltheart, & Rastle, 2016; Rastle & Coltheart, 2000), suggesting an important role for morphology in stress assignment.

Other variables that received little or no attention are orthographic consonant-vowel (CV) structure (i.e., the sequence of consonants and vowels composing a word; e.g., the orthographic CV-structure of *macchia* is CVCCCVV) and number of syllables. Whether the association between stress patterns and CV-structures in the lexicon may affect stress assignment is an open issue. The same is true for the possible role played by the number of syllables. One may ask, e.g., whether words differ in the distribution of stress patterns according to the number of their syllables, and how this information interacts with other sources for stress. In particular, a still unanswered question concerns which words should be considered when computing stress neighborhood. While so far stress neighborhood has been computed by counting all words sharing an ending regardless of the number of their syllables, it is still unclear whether words with a different number of syllables may give a different contribution to stress neighborhood.

To sum up, research focused on stress dominance and stress neighborhood, considering them the main factors affecting stress assignment in Italian. However, we still do not know whether and to what extent other properties of words also work as cues to stress assignment. Good candidates include grammatical category, word beginnings and endings of different length, CV-structures, and number of syllables.

Resources for investigating Italian stress

The lack of research on the role of the properties discussed above in stress assignment may be due to the inadequacy of current databases. Since Colombo's (1992) seminal study, investigations addressing stress assignment in Italian have relied on three major frequency dictionaries: The dictionary by Bortolini, Tagliavini, and Zampolli (1971), the *Barcelona corpus* (Istituto di Linguistica Computazionale, 1989), and *CoLFIS*, a more recent corpus and frequency count of more than 3 million written Italian word occurrences (Bertinetto *et al*., 2005). Since no one of the above databases included stress information, researchers annotated stress information manually. Thus, those resources were unsuited to give a picture of how stress patterns distribute in Italian. At this date, the only study that provides such information is *BDVDB: A database for the Italian basic dictionary* (Thornton,

Burani, & Iacobini, 1997). According to Thornton and collaborators (1997), most Italian words with more than two syllables have penultimate stress (85.3%), some have antepenultimate stress (12.7%), and very few have ultimate stress (1.8%) (note 3). The same is true when number of syllables and grammatical category are taken into account: Penultimate stress is always the dominant pattern, with a minority of antepenultimate stress forms. However, some doubts can be cast upon such data. *BDVDB* is based on just the 5,000 most frequent (plus the 2,000 most "available") Italian words, collected by De Mauro (1991) as a list of lemmas. As a result, information is provided for neither less frequent words nor word forms in general. Note that in Italian each lemma may correspond to several word forms, and that this is especially true for verbs, which show different inflectional markers for each of the twelve simple tenses, each of the three conjugations, and each person. Most importantly, verbs show alternating stress patterns and number of syllables within inflectional paradigms. Thus, counts made on lemmas may differ dramatically from those made on word forms. Besides, nouns are possibly given too much weight in *BDVDB*, since they cover the vast majority of words in De Mauro's (1991) list (4,580 on 7,000).

Another major limitation in current research is the lack of a child database providing stress information. Indeed, the distribution of stress patterns in the written materials children are exposed to may have different stress properties with respect to adult materials. So far, however, statistics for studies conducted on children were drawn from *BDVDB*, and the *Lessico elementare* (*Elementary Lexicon,* Marconi, Ott, Pesenti, Ratti, & Tavella, 1993) served as the main source for experimental materials. Being built on both child-directed texts and texts written by children, the latter database is arguably the most appropriate resource for children. However, it does not provide any phonological information, and thus, researchers had to annotate stress manually. As such, longitudinal studies on the development of stress assignment have been fairly impossible to date.

The present work aims to make up for these and other limitations. In the following, we describe *Q2Stress* (Cue-To-Stress), a lexical database providing information for multiple variables of interest for research on stress assignment in Italian. As the name suggests, *Q2Stress* relies on the claim that several sources of distributional information may act as cues to stress assignment, including not

only stress dominance and stress neighborhood, but also variables for which no empirical evidence is currently available for Italian, such as grammatical category, word beginnings and endings of different length, CV-structure, and number of syllables. *Q2Stress* will thus be a powerful tool to shed light on issues such as 1) which are the independent roles of the different variables in assigning stress, 2) how these variables may interact, and 3) which cross-linguistic differences and similarities are characteristic of stress assignment in Italian. In addition, *Q2Stress* includes data coming from both child-directed and adult corpora, thus allowing researchers to examine to what extent differences in performance between children and adults are due to either the properties of their lexicons or the development of cognitive skills. Finally, being derived from large corpora of word forms (rather than lemmas), *Q2Stress* is not affected by the drawbacks of Thornton and collaborators (1997).

Methodology

Q2Stress has been built up on the basis of both an adult and a child-directed corpus. The basis for the first (adult) part of *Q2Stress* is *phonItalia* (Goslin, Galluzzi, & Romani, 2014), an open access lexical database providing orthographic and phonological information for 120,000 word forms extracted from *CoLFIS* (Bertinetto et al., 2005). The choice of *phonItalia* instead of *CoLFIS* was motivated by the fact that the former, but not the latter, contains information about the stress pattern of word forms. *Q2Stress* is based on *phonItalia version 1.1* (http://www.phonitalia.org) and derives all of its fields from *phonItalia*'s main lexicon ("phonItalia 1.10 – word forms.txt").

The second part of *Q2Stress* derives from *Lessico elementare [Elementary lexicon]* (Marconi et al., 1993), a frequency count of about 1,000,000 occurrences based on materials equally divided between those written by adults for children (age range: 6-10 years) and those written by children themselves. Since our focus is on how reading, rather than writing, is affected by distributional knowledge about stress, we performed our analyses only on the child-directed half (about 25,000 word forms). For each word form, the corpus included information concerning frequency, lemma, and grammatical category. Grammatical category markers were converted to *phonItalia*'s codification in the first place. Then, stress information, as well as phonological representations, number of syllables,

syllabic boundaries, number of letters and phonemes, and orthographic and phonological CV-structure, were obtained from *phonItalia* by comparing the two databases. A remarkable proportion of words included in the child-directed corpus (over 20%) did not occur in the larger adult corpus, which suggests that the two databases differ considerably. As we shall see, the child-directed corpus shows a relatively higher number of verbs than the adult corpus. An analysis of the child-exclusive forms reveals that about half are verbs, especially in enclitic form (e.g., *acchiapparci* 'to catch us'). The remaining forms are typically found in texts for children, such as diminutive and augmentative nouns and adjectives (e.g., *regalini* 'little gifts'), and interjections. Information for those forms was obtained using the phonological transcription module from the Italian *Festival* text-to-speech system (Cosi, Gretter, & Tesser, 2001), which is the same software used by Goslin and collaborators (2014). The output of the software was manually checked. Cases for which stress position is uncertain in Italian (e.g., RUbrica vs. ruBRIca 'address book') were assigned the first stress pattern listed in the De Mauro Italian Dictionary (De Mauro, 2000).

Note that the purpose of the phonological transcription process was to obtain the stress pattern and number of syllables for each word form. Phonological representations (e.g., /ˈkɔsto/ 'cost'), though available, were not used. Although Italian is a transparent language, phonological representations are sometimes less ambiguous with respect to stress position than orthographic representations are. For example, in such Italian varieties as Tuscan and Roman, open-mid vowels (e.g., /ɔ/) are more predictive of stress position than close-mid vowels (e.g., /o/) are, but in print, they correspond to the same grapheme (i.e., <o>). As *Q2Stress* aims to model the Italian reader's knowledge about stress when he/she is engaged in reading aloud, only orthographic units shall be used, thus maintaining the ambiguity of Italian spelling with respect to stress position.

Monosyllabic words were excluded from the analysis of both databases, as they are of no interest for stress assignment (at least in single-word reading). Although stress assignment to disyllabic stimuli is straightforward in Italian (i.e., they have penultimate stress unless a stress diacritic marks the last syllable), disyllables were included since they may contribute to the overall

distribution of stress patterns, e.g., by enhancing a bias towards the dominant pattern. Finally, all words with three or more syllables were included.

Analyses were made on both word types and tokens (note 4). Among types, homographs and words occurring in idioms were counted just once. Thus, type analyses were made only on non-identical word forms, which amount to 94,051 in *phonItalia* and 22,081 in *Lessico elementare*. Tokens, on the other hand, amount to 2,179,735 in *phonItalia* and 272,754 in *Lessico elementare*.

In both the adult and the child-directed parts of the database, we considered the following variables of interest for stress assignment: 1) number of syllables; 2) grammatical category; 3) word beginnings (from 1 to 4 letters long); 4) word endings (from 1 to 4 letters long); 5) word endings' CV-structure (3-letter endings only); 6) whole-word CV-structure. Unfortunately, because *phonItalia* does not include morphological information, it was not possible to investigate the role of affixes on stress assignment at this time.

A few remarks should be made concerning the variables 3), 4), and 5). With regard to word beginnings, since no criteria have been developed until now, we chose to consider word-initial units of increasing length, from one up to four letters. Corpus analyses show that longer beginnings are better predictors of stress position than shorter ones (Monaghan et al., 2016), however future research will shed light on which size best accommodates results from empirical investigations.

Being the focus of several investigations, word endings are more clearly defined. In Italian, stress neighborhood refers to word endings starting from the nucleus of the penultimate syllable. According to this definition, word endings may vary in length. However, since three-letter endings (with a VCV structure, e.g., ab-ete 'fir') are the most frequent in Italian, almost the totality of the literature is concerned with them. Endings may be simply defined in terms of their length, though (e.g., Monaghan et al., 2016). Thus, it might be the case that endings as short as one or two letters provide a sufficient account of how stress is assigned in Italian. Alternatively, readers might use units as long as three or even four letters. As an approach to answering this empirical question, and mirroring the data for word beginnings, we provided data for endings ranging from 1 to 4 letters,

regardless of the position of the nucleus of the penultimate syllable. We decided, however, to restrict information concerning word endings' CV-structure to three-letter endings only. This was done in the interest of simplicity, as inspection of word endings' CV-structures revealed that three-letter endings are by far the most informative in terms of stress pattern distribution. Accordingly, we will spend more time commenting on three-letter endings relative to other endings in the following sections.

The database

Q2Stress is freely available at http://www.istc.cnr.it/grouppage/databases. Being thought as a device for researchers, *Q2Stress* comes with two tools subserving two main research purposes: a set of functions for exploration and selection of stimuli (see the "scripts" folder), and a set of Summary Tables for overall data analysis (see the "summary tables" folder).

The functions are written in R (https://www.r-project.org/) and can be easily used also by R neophytes. The functions allow the user to specify a word beginning (e.g., *to*-) or a word ending (e.g., *-ola*) and a stress pattern (e.g., penultimate stress), and obtain all words starting with that beginning and bearing the specified stress pattern (e.g., *torrone* 'nougat', *tovaglia* 'tablecloth', etc.). More detailed queries taking into account other parameters (e.g., the number of syllables, grammatical category) can also be done. Note that the output of each query is saved in a txt file. Instructions to use the functions can be found in the Supplementary materials.

The summaries of the overall type and token counts for word beginnings, word endings, and CV-structures can be found in the tables, which report data for adults and children separately. Summary tables report the percentage and the number of words with each type of stress pattern (ultimate, penultimate, antepenultimate, or preantepenultimate) in which each unit (beginning, ending, or CV-structure) appears. For example, they show that in the adult type count the word ending *-ero* appears in 24 penultimate stress words and 220 antepenultimate stress words, corresponding to 9.8% and 90.2% of all words ending in *-ero*, respectively. More detailed information about summary tables can be found in the Supplementary materials. In what follows, we report some of the main findings resulting from these tables.

Descriptive analyses for the adult frequency count

We report descriptive analyses for *phonItalia* 1.1. The variables we focus on distribute as follows. Three-syllable and four-syllable words are the most frequent in terms of types (31.3% and 35.5%, respectively; disyllables: 11.4%), but are outweighed by disyllables in terms of tokens (disyllables: 44.5%; three-syllables: 31.6%; four-syllables: 16.7%). Verbs, nouns, and adjectives are the most frequent grammatical categories when either types (35.8%, 28.5% and 20.1%, respectively) or tokens (25.8%, 33.7%, and 14.7%, respectively) are considered. Finally, VCV word endings cover more than half of the types (56.9%) and half of the tokens (49.6%) of the database, followed by CCV (30.8% of types, 36.2% of tokens) and CVV word endings (6.4% of types, 9.6% of tokens). Data from word beginnings and endings, as well as whole-word CV-structures, are more sparse.

Overall distribution. As shown in Table 1, our data confirm the picture provided by Thornton and collaborators (1997) on lemmas, with the frequency order of Italian stress patterns being 1) penultimate, 2) antepenultimate, and 3) ultimate (note 5).

-Table 1-

This holds true when both types and tokens are taken into account, with a slightly stronger penultimate-stress dominance in token than in type percentages (87% vs. 76.7%): This difference is driven by disyllables, which account for 44.9% of tokens and bear penultimate stress most of the time. When disyllables are excluded, the difference between types (penultimate: 75.9%; antepenultimate: 20%; ultimate: 3.8%) and tokens (penultimate: 79.9%; antepenultimate: 17.3%; ultimate: 2.6%) is smaller. Note that type measures better fit results from studies addressing stress assignment: In a reading aloud task Burani and Arduino (2004, Experiment 2) manipulated stress neighbors in terms of types while controlling them in terms of tokens and found a larger stress neighborhood effect for word endings with higher types (see also Sulpizio et al., 2013). Accordingly, in what follows, only type counts will be discussed. The following tables and figures, however, will report measures for both types and tokens.

Number of syllables and Grammatical category. Penultimate stress is the most frequent pattern even when one considers different numbers of syllables and grammatical categories. However, the higher the number of syllables in a word, the more likely the word is to have an antepenultimate stress pattern (Figure 1).

-Figure 1-

As far as grammatical category is concerned, we found (Figure 2) that, although penultimate stress is the dominant pattern for each grammatical category in Italian, verbs have a lower proportion of penultimate stress words as compared to nouns (70.3% *vs*. 84.9%). Adjectives stand half-way between nouns and verbs (76.2%), while adverbs are the most biased towards penultimate stress (95.3%). In sum, Italian shows some variability in the distribution of stress patterns within grammatical categories; however, while languages like English show a preference for the dominant (first-syllable) stress pattern in disyllabic nouns and a preference for the nondominant (second-syllable) stress pattern in disyllabic verbs (Arciuli & Cupples, 2006; 2007), the dominant pattern is always preferred in all grammatical categories in Italian. Besides, all categories except adverbs are in line with the overall tendency of a decrease of penultimate stress patterns with the increase of number of syllables.

-Figure 2-

Word beginnings. We extracted word beginnings of one up to four letters. Almost all of one-letter and two-letter beginnings mirror the overall stress pattern distribution (i.e., they are associated with the penultimate stress pattern). Exceptions are few (less than 5% of all beginnings), cover few word forms, and are not strongly associated with any other stress pattern. Longer beginnings vary to a larger extent than shorter ones, as they include beginnings showing no bias towards penultimate stress (e.g., *acr-*, *apos-* appear more often in antepenultimate stress words) all the way up to beginnings appearing always in penultimate stress words (e.g., *paz-*).

Besides percentages, it is important to consider also the total number of words each beginning appears in. In fact, while some beginnings are very frequent (e.g., the first five one-letter beginnings

appear in more than half of the word forms), others are not. Since numerosity (i.e., the raw number of words having the same cue) was found to play a role for word endings (see below), this may hold true also for beginnings. Accordingly, high-frequency beginnings may be better cues to stress than low-frequency ones. However, overall, beginnings seem to be less informative of stress position than endings, confirming the analyses reported by Monaghan and collaborators (2016).

Word endings. The most common one-letter and two-letter endings (e.g., *-a*, *-i*, *-re*, *-to*) are strongly associated with penultimate stress, and thus appear to offer few cues to stress position. Deviations from this trend concern endings that are pretty rare in Italian, such as those appearing mostly in French and English loanwords (e.g., *-l*, *-t*, *-er*, *-in*).

For three-letter endings, we focused on the words that bear no stress diacritic. These included VCV, CCV, CVV, CVC, and VVV endings. The distribution of stress patterns for such CV-ending structures is reported in Figure 3. VCV and CCV are the most frequent endings, accounting for 56.9% and 30.8% of all word forms, respectively. Although penultimate stress is the dominant pattern for both endings, CCV endings show a higher proportion of penultimate stress words (94.1%) than VCV ones (71.4%). In fact, CCV endings include a large proportion of words that bear penultimate stress because of the phonological rule (i.e., they have a closed penultimate syllable, e.g., *conCERto* 'concert'); they also include a small proportion of words with an open penultimate syllable (i.e., not ending in a consonant) with a stop + liquid onset in the last syllable,, which can bear any stress pattern (cf. *puLEdro* 'colt' and *VERtebra* 'vertebra'). Statistics for CVV and VVV endings are similar to CCV endings. In contrast, CVC endings are peculiar, as they show a tendency towards ultimate stress. The reason for this is that CVC is not a regular Italian ending, and most words with such ending are either English or French loanwords.

-Figure 3-

Because of their low frequency, we will not go into details of other CV-structures. Note, however, that penultimate stress is the dominant pattern for most of them as well (for details, see types.txt and tokens.txt in the CV-ending folder). As it might be expected, we found an overwhelming

number (325) of different VCV endings, which account for over 90% of all word forms for which stress cannot be derived by rule. Like beginnings, VCV endings vary in both the total number (numerosity) of words they appear in and the proportion of penultimate and antepenultimate stress patterns those words have. Both measures were found to have a role in stress assignment in Italian (Burani & Arduino, 2004; Burani et al., 2014; Sulpizio et al., 2013): In nonword reading, for example, word endings with many stress friends (i.e., words sharing the ending and stress pattern) received more responses consistent with stress neighborhood than word endings with the same percentages of stress friends but lower numerosity. Sulpizio and collaborators (2013) argue that stress neighborhood effects are to be found only for endings with both a large proportion and number of stress friends. In contrast, both small-sized and balanced neighborhoods are ambivalent, since they provide no clear cue for stress (Colombo et al., 2014). Q2Stress lists word endings of both classes. Most endings, such as -*ofo* and -*abo*, have low to medium frequency, and are thus closer to ambivalence. The same is true for endings, such as -*isi* and -*ene*, that have good-sized neighborhoods but show an almost equal proportion of penultimate and antepenultimate stress friends. The most part of word forms, however, is covered by endings that do have large neighborhoods and are clearly biased towards either penultimate (e.g., -*ore*, -*ari*) or antepenultimate stress (e.g., -*olo*, -*ico*), with the former outweighing the latter.

Four-letter endings confirm the tendency of many endings associated with few word forms and few endings associated with many word forms, either mostly stressed on the penultimate or mostly stressed on the antepenultimate syllable. Interestingly, there are remarkable differences between four-letter endings that share the last three letters: For example, -*tina* is strongly associated with penultimate stress (97.5%), whereas for -*mina* penultimate and antepenultimate patterns are pretty balanced (about 50%). Although this is not always the case, such occurrences suggest that four-letter endings may be better predictors of stress position than three-letter endings are.

When one takes number of syllables and grammatical category into account the picture becomes less clear-cut. As noted above, there is a tendency for words with increasing number of syllables to bear antepenultimate stress more often. This tendency is clearly visible in the most

common one-letter endings. However, the situation changes as longer endings are considered. While some two-letter endings show a steep increase in the proportion of antepenultimate stress words as the word gets longer (e.g., *-ci*, *-he*), others show no sign of this tendency (e.g., *-ni*, *-va*). Similarly, one can distinguish three-letter and four-letter endings for which stress is and is not attracted backwards as the word length increases. Overall, the distribution of stress patterns for most endings associated with a large number of word forms changes minimally with increasing number of syllables, but a few endings (e.g., *-ile*, *-done*) have different behaviors. For instance, *-ile* has, overall, an antepenultimate stress neighborhood; however, three-syllable words ending in *-ile* show pretty balanced proportions of penultimate and antepenultimate stress words. Readers might be sensitive to such item-specific information. Indeed, Sulpizio and collaborators (2013, Experiment 2) failed to report a stress neighborhood effect for three-syllable nonwords ending in *-ile* (note 6).

For the sake of simplicity, we restrict discussion of grammatical category to three-letter endings with a VCV structure. Several such endings show remarkable differences in how stress patterns distribute within grammatical categories. Verbs, in particular, behave differently from other categories. For instance, consider *-era,* a word ending with a prevalently penultimate stress neighborhood. While nouns and adjectives ending in *-era* (e.g., *panTEra* 'panther', *alTEra* 'proud') show a preference for penultimate stress, the opposite is true for verbs (e.g., *conSIdera* 'he/she/it considers'), which bear antepenultimate stress most of the time. Thus, while grammatical categories *per se* have little impact on the distribution of stress patterns at the general level, they seem to have a role at the item-specific level, that is when word endings are considered.

CV-structures. Almost all CV-structures are biased towards penultimate stress. Variability among CV-structures is mainly accounted for by the structure of the ending, with CCV endings showing a higher proportion of penultimate stress words than VCV and other endings. However, if the analysis is restricted to CV-structures with the same ending (e.g., VCV), some noteworthy differences remain (e.g., between VCVCV and CVCCCVCV, that bear penultimate stress 62.5% and 83% of the times, respectively).

Descriptive analyses for the child-directed frequency count

The distributional properties of the child-directed part of *Q2Stress* mirror those of the adult part to a large extent. Three-syllable and four-syllable words are more frequent than disyllables in terms of types (disyllables: 15.8%; three-syllables: 38.1%; four-syllables: 33.8%), but are outweighed by disyllables in terms of tokens (disyllables: 50.6%; three-syllables: 32.9%; four-syllables: 12.8%). Similarly to the adult database, verbs, nouns, and adjectives are the most frequent grammatical categories in terms of both types and tokens. However, verbs have dramatically more types than any other category (59.9% of all word types, that is twice as many as nouns). Despite being numerous, verb forms tend to occur with a low frequency, being second to nouns in terms of tokens, with similar proportions to the adult database (29.8% *vs.* 37.8%, respectively). Counts for ending CV-structures also mirror the adult part pretty much: VCV endings cover more than half of the types (56.1%) and almost half of the tokens (48.5%), followed by CCV (32.3% of types, 35% of tokens) and CVV endings (4.9% of types, 9.6% of tokens).

Overall distribution. Table 2 shows that also for children penultimate stress is the dominant pattern , followed by antepenultimate and ultimate stress. Given the high degree of similarity between adult and child-directed data, in what follows we will discuss only the cases in which the latter markedly differ from the former. As before, only types will be considered.

-Table 2-

Number of syllables and Grammatical category. Although penultimate stress is the dominant pattern, it shows lower percentages with increasing number of syllables, with a more pronounced tendency than what was found in the adult database. As shown in Figure 4, words with as many as 5 syllables have a weak bias towards penultimate stress, and polysyllables with more than 6 syllables even show the reversed pattern, i.e., with a preference for antepenultimate stress (at least when types are considered) (note 7).

-Figure 4-

No reversion in the preference for the penultimate and antepenultimate stress patterns is found when one considers the distribution of stress patterns across grammatical categories (Figure 5). However, one can notice a remarkable difference between nouns (which bear penultimate stress 88.9% of the time) and verbs (which bear penultimate stress only 67.4% of the times).

This difference is more relevant when both number of syllables and grammatical category are considered. While nouns have a steady proportion of penultimate stress words with increasing number of syllables (e.g., the 84.40% of 3-syllabic nouns and the 85.11% of 6-syllabic nouns bear penultimate stress), the proportion of penultimate stress verbs is progressively reduced (e.g., while 3-syllabic verbs bear penultimate stress 74.63% of the time, the percentage drops to 65.15% for 4-syllable and 50.68% for 5-syllable verbs). Adjectives and adverbs follow the same pattern shown in the adult data.

-Figure 5-

Word beginnings. As for adults, while most beginnings are associated with a majority of penultimate stress words, longer beginnings show more variability than shorter ones. The heterogeneity among beginnings increases when the number of syllables is considered, with beginnings appearing in longer words being less strongly associated with penultimate stress than beginnings appearing in shorter words . Finally, beginnings also vary in numerosity, with some beginnings appearing in a relatively large number of words (e.g., *sco*-, *per*-) and some appearing in fewer words (e.g., *sus*-, *lit*-).

Word endings. As for the adult part, one-letter and two-letter endings are strongly associated with penultimate stress. Also three- and four-letter endings mirror the patterns reported for adults, with. many endings associated with few word forms and few endings associated with many word forms; in the latter case, there is often a majority of either penultimate or, more rarely, antepenultimate stress words.

-Figure 6-Results relative to number of syllables and grammatical category mirror what was found for the adult part.

Concerning the impact of the former, inspection of longer endings (i.e., three and four letters) revealed that only some endings (e.g., *-ene*, *-ersi*) show a tendency towards antepenultimate stress as words get longer. . Focusing on three-letter endings only, this variability within endings is more common when grammatical categories are considered: Several word endings show a reversed, or remarkably different, distribution of penultimate and antepenultimate stress patterns according to the grammatical category of the words they appear in.

CV-structures. Overall, no effect of CV-structures on the distribution of stress patterns is readily apparent for three-syllable words, which are associated with penultimate stress most of the time. A look at four-syllable words, however, reveals that several structures with VCV endings are stress-neutral (with about 50% of penultimate stress words, e.g., CVCCVCCVCV), whereas others show a (moderate) bias towards penultimate stress (with more than 60% of penultimate stress words, e.g., VCCVCCVCV). The same is true for words with five or more syllables, which also have CV-structures biased towards antepenultimate stress.

Discussion

We have presented *Q2Stress*, a database that provides information for multiple cues to stress assignment in reading. The present database permits to expand opportunities of research in reading in Italian in at least two directions.

First, distributional information other than stress dominance and stress neighborhood is available in the lexicon. Our data offer the opportunity to inquire the role of such information and investigate how it interacts with stress dominance and stress neighborhood. By querying our database, we have found that penultimate stress words decrease with increasing number of syllables, and nouns and adverbs include relatively more penultimate stress words than verbs and adjectives. Number of syllables and grammatical information seem especially important in relation to stress neighborhood, as in some cases stress patterns for a given ending distribute differently according to which polysyllable and/or grammatical category are considered. In addition, word beginnings, like word

endings, vary in the number of words they appear in, and especially longer beginnings may be either stress-neutral or biased towards penultimate or antepenultimate stress. Finally, specific CV-structures, though less informative, show some instances of weak *vs.* strong biases towards dominant stress.

We believe that, besides providing key insights concerning the distributional knowledge used by readers and the computational levels involved, each of the above findings may be conducive to a better understanding of the stress assignment process within the reading system: Looking at how stress assignment may be affected by sublexical and lexical information – such as orthographic cues and grammatical category, respectively – can shed further light on both the time dynamic of the lexical and sublexical computations and the mechanisms of interaction between different reading routines.

Secondly, for the first time, *Q2Stress* makes available a resource for investigating stress assignment in children. The child-directed part of the database may, in fact, provide an answer to several questions concerning which information is available to children and how it differs from that for adults. One might ask, for instance, 1) whether the stress dominance effect found in young children is partially due to a higher exposure to penultimate stress words, and 2) whether child data show the same properties of stress neighborhood that were found for adult data. Our data suggest that children as compared to adults are apparently not exposed to enhanced stress dominance in written texts. We found that, overall, the proportion of penultimate stress words is about the same in the child and the adult data. With respect to stress neighborhood, word endings range from small-sized to large-sized stress neighborhoods, and larger neighborhoods are generally biased towards either penultimate or antepenultimate stress. Thus, similar to what was found for adults, in the texts written for children there are both ambivalent word endings and endings that are good candidates to act as cues to stress. When compared to word endings extracted from the adult database, the main differences concern only small-sized neighborhoods. Note, however, that some medium- and high-frequency endings are more common in the child than in the adult data (e.g., *-ino*, *-ere*) and, conversely, some are more common in the adult than in the child-directed data (e.g., *-ico*, *-ile*). These differences should be taken into account when constructing experimental materials for developmental studies.

In the following, we discuss three of the multiple pathways that the new database may contribute to follow for future research. *Q2Stress* is intended to promote investigations addressing the relationship between stress and various sources of information, most of which have received little attention so far. This is especially true for information concerning the grammatical category of words. Current models of reading (e.g., DRC: Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; CDP++: Perry et al., 2010) include no information about grammatical category. However, some empirical studies showed that grammatical category may be a cue to stress. In both English (Arciuli & Cupples, 2006) and Russian (Jouravlev & Lupker, 2014), an interaction was found between grammatical category and stress neighborhood. Arciuli and Cupples (2006) argued that their results support the claim that orthographic units cue simultaneously the grammatical category and the stress pattern they are typically associated with. Thus, in accordance with a constraint satisfaction process, an advantage is found only when reading, e.g., a first syllable stress noun with an ending that occurs in a majority of nouns and a majority of first syllable stress words. However, in both English and Russian stress is conditioned by grammatical category (i.e., different categories show different stress pattern distributions). Thus, the alternative explanation that orthographic units cue a given grammatical category, which in turn cues its typical stress pattern, cannot be rejected. Thanks to its properties, Italian can contribute to solve the issue. While penultimate stress is the dominant pattern in every grammatical category of Italian, we found that several stress neighborhoods show inconsistencies across categories. For example, most verbs ending in *-era* bear antepenultimate stress although *-era* has a penultimate stress neighborhood and other categories show the same tendency. If readers were found to use such specific information, there would be convincing evidence that grammatical categories do not cue their typical stress pattern. In fact, there is no typical stress pattern for Italian grammatical categories – they all prefer penultimate stress. This line of research may shed light on the mechanisms involved in stress neighborhood and, more in general, on the relation between prosodic and grammatical information (for a first attempt in Italian, see Spinelli, Sulpizio, Primativo, & Burani, 2016). In addition, our data show that not only grammatical category but also the number of syllables affect how stress patterns distribute within words sharing an ending. While current research assumes

that stress neighborhood depends on the total number of stress friends, it may be the case that more specific distributional information is used by readers.

Orthographic cues to stress may take place also in word-initial position. Arciuli and collaborators (Arciuli & Cupples, 2007; Arciuli et al., 2010) found that English beginnings, though less strongly than endings, affect stress and grammatical category assignment (see also Kelly, 2004). Corpus analyses by Monaghan and collaborators (2016) also show that word beginnings are good predictors of stress position in several languages, including Italian. To our knowledge, however, no investigation has empirically addressed this issue for Italian. While stress neighborhood seems to account for the overall performance in reading aloud, the variability among specific stimuli having the same ending is not readily explained. Beginnings are the most likely to have a role in this. Q2Stress now provides plenty of information to answer this question. We made available information concerning the distribution of stress patterns for beginnings of increasing length, from one to four letters. Whether orthographic beginnings are used by readers, and which grain-size units are used, is an empirical matter. Our analyses, however, indicated that longer beginnings (i.e., 3-4 letters) provide a better account of the variability in the lexicon, and may thus act as more effective cues to stress. Another finding was that, for some beginnings, the distribution of stress patterns varies according to the number of syllables. As a result, the number of syllables should also be taken into account when investigating orthographic beginnings.

The availability of multiple potential cues for stress assignment paves the way for the investigation of the relative strength of these cues and their relations. So far, research on stress assignment has been investigating some of the possible cues to stress without considering them together (for a recent attempt in this direction, see Mousikou, Sadat, Lucas, & Rastle, in press). Future research, however, may overcome this limit by simultaneously investigating multiple cues and exploring whether they equally contribute to stress assignment or their relevance is hierarchically arranged instead. Investigating the existence of a hierarchy among cue predictors to stress will help to shed further light on both the organization of lexical and sublexical knowledge in the reading system, and the relation these types of knowledge have with the word's prosodic properties.

In sum, much information in *Q2Stress* can be useful to investigate which and how distributional knowledge affects stress assignment in Italian. The variables we considered here are not the only ones that may play a role in stress assignment, though. For example, in several languages, including Italian, stress is morphologically conditioned (i.e., affixes are either stressed or unstressed; see, for English, Rastle & Coltheart, 2000), and there is evidence from Greek that phonetic similarity also plays a role (Protopapas et al., 2007). Neither of these sources of information was included in *Q2Stress*. Yet another possibility is to study the relation between syllables and stress: In Italian, some syllables may appear in stressed form more often than other syllables do, and this may have an impact on the easiness with which they are assigned stress. Thus, further expansions are needed for a full picture of the stress information available to readers.

Conclusion

Q2Stress was aimed to fill the lack of resources for researchers interested in stress processing in reading Italian. It was designed as a user-friendly resource, as it comes with both tables for overall data analysis and scripts allowing users to query the database using multiple criteria. Most importantly, it was intended to support and promote research in the field by including both variables known to affect stress assignment in Italian and variables for which evidence is still lacking. Our analysis revealed that many of those may play a role. However, it is the job of empirical research to establish whether readers actually use them and how this affects our view of the reading system.

Acknowledgment: Special thanks go to Lucia Marconi (ILC-CNR) for extracting word form information from the *Lessico Elementare* and allowing us to use and share it.

References

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *The Quarterly Journal of Experimental Psychology*, *59*, 920-948.

Arciuli, J., & Cupples, L. (2007). Would you rather "embert a cudsert" or "cudsert an embert"? How spelling patterns at the beginning of English disyllables can cue grammatical category. In A. Schalley & D. Khlentzos (Eds.), *Mental states: Language and cognitive structure* (Vol. 2, pp. 213–238). Amsterdam, the Netherlands: John Benjamins.

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language*, *63*, 180-196.

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

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., & Thornton, A. M. (2005). *CoLFIS (Corpus e Lessico di Frequenza dell'Italiano Scritto) [Corpus and Frequency Lexicon of Written Italian]*. Institute of Cognitive Sciences and Technologies, CNR, http://www.istc.cnr.it/grouppage/colfisEng

Bortolini, U., Tagliavini, C., & Zampolli, A. (1971), *Lessico di frequenza della lingua italiana contemporanea [A frequency count of contemporary Italian language]*. Milan, Italy: Garzanti.

Brown, P., Lupker, S. J., & Colombo, L. (1994). Interacting sources of information in word naming: A study of individual differences. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 537-554.

Burani, C., & Arduino, L. S. (2004). Stress regularity or consistency? Reading aloud Italian polysyllables with different stress patterns. *Brain and Language*, *90*, 318-325.

Burani, C., Paizi, D., & Sulpizio, S. (2014). Stress assignment in reading Italian: Friendship outweighs dominance. *Memory & Cognition*, *42*, 662-675.

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance, 18, 987-1003.*

Colombo, L., Deguchi, C., & Boureux, M. (2014). Stress priming and statistical learning in Italian nonword reading: evidence from children. *Reading and Writing*, *27*, 923-943.

Colombo, L., & Sulpizio, S. (2015). When orthography is not enough: The effect of lexical stress in lexical decision. *Memory & Cognition*, *43*, 811-824.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204-256.

Cosi, P., Gretter, R., & Tesser, F. (2001). Festival parla italiano! [Festival speaks Italian]. In E. Magno Caldognetto & P. Cosi (Eds.), *Proceedings of GFS2000, XI Giornate del Gruppo di Fonetica Sperimentale*. Padova, Italy: UNIPRESS.

Cutler, A. & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*, 133-142.

De Mauro, T. (1991). *Guida all'uso delle parole* [*A guide for the use of words*] *(11th ed.)*. Rome, Italy: Editori Riuniti.

De Mauro, T. (2000). *Il dizionario della lingua italiana* [*The dictionary of Italian language*]. Turin, Italy: Paravia.

Goslin, J., Galluzzi, C., & Romani, C. (2014). PhonItalia: a phonological lexicon for Italian. *Behavior Research Methods, 46*, 872-886. http://www.phonitalia.org/

Howard, D., & Smith, K. (2002). The effects of lexical stress in aphasic word production. *Aphasiology*, *16*, 198-237.

Istituto di Linguistica Computazionale del CNR di Pisa (1989). *Corpus di Italiano scritto contemporaneo [A corpus of contemporary written Italian*]. Unpublished manuscript.

Jarmulowicz, L., Taran, V. L., & Hay, S. E. (2008). Lexical frequency and third-graders' stress accuracy in derived English word production. *Applied Psycholinguistics*, *29*, 213-235.

Jouravlev, O., & Lupker, S. J. (2014). Stress consistency and stress regularity effects in Russian. *Language, Cognition and Neuroscience*, *29*, 605-619.

Kelly, M. H. (1988). Phonological biases in grammatical category shifts. *Journal of Memory and Language*, *27,* 343–358.

Kelly, M. H. (2004). Word onset patterns and lexical stress in English. *Journal of Memory and Language, 50*, 231-244.

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 389-403.

Ktori, M., Tree, J. J., Mousikou, P., Coltheart, M., & Rastle, K. (2016). Prefixes repel stress in reading aloud: Evidence from surface dyslexia. *Cortex*, *74*, 191-205.

Laudanna, A., Thornton, A. M., Brown, G., Burani, C., & Marconi, L. (1995). Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente [A new corpus of contemporary written Italian]. In S. Bolasco, L. Lebart, & A. Salem (Eds), *III Giornate internazionali di Analisi Statistica dei Dati Testuali* [*3rd International Conference on Statistical Analysis of textual data*] Volume I (pp. 103-109). Roma, Italy: Cisu.

Marconi, L., Ott, M., Pesenti, E., Ratti, D., & Tavella, M. (1993). *Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementari [Elementary lexicon: Statistical data for Italian written and read by elementary school children]*. Bologna, Italy: Zanichelli.

Monaghan, P., Arciuli, J., & Seva, N. (2016). Cross-linguistic evidence for probabilistic orthographic cues to lexical stress. In J. Thomson & L. Jarmulowicz (Eds.), *Linguistic rhythm and literacy.*

Trends in Language Acquisition Research series (TiLAR) (pp. 215-236). Amsterdam, The Netherlands: John Benjamins.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*, 43–71.

Mousikou, P., Sadat, S., Lucas, J., & Rastle, K. (in press). Moving beyond the monosyllable in models of skilled reading: mega-study of disyllabic nonword reading. *Journal of Memory and Language.*

Pagliuca, G., & Monaghan, P. (2010). Discovering large grain sizes in a transparent orthography: Insights from a connectionist model of Italian word naming. *European Journal of Cognitive Psychology*, *22*, 813–835.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*, 106–151.

Protopapas, A., Gerakaki, S., & Alexandri, S. (2006). Lexical and default stress assignment in reading Greek. *Journal of Research in Reading*, *29*, 418–432.

Protopapas, A., Gerakaki, S., & Alexandri, S. (2007). Sources of information for stress assignment in reading Greek. *Applied Psycholinguistics*, *28*, 695-720.

Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language*, *42*, 342-364.

Sereno, J. A. (1986). Stress pattern differentiation of form class in English. *The Journal of the Acoustical Society of America*, *79*, S36.

Smith, P., Baker, R., & Groat, A. (1982). Spelling as a source of information about children's linguistic knowledge. *British Journal of Psychology*, *73*, 339–350.

Spinelli, G., Sulpizio, S., Primativo, S., & Burani, C. (2016). Stress in context: morpho-syntactic properties affect lexical stress assignment in reading aloud. *Frontiers in Psychology, Language Science, 7*: 942.

Sulpizio, S., Arduino, L. S., Paizi, D., & Burani, C. (2013). Stress assignment in reading Italian polysyllabic pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 51-68.

Sulpizio, S., Burani, C., & Colombo, L. (2015). The process of stress assignment in reading aloud: Critical issues from studies on Italian. *Scientific Studies of Reading*, *19*, 5-20.

Sulpizio, S., & Colombo, L. (2013). Lexical stress, frequency, and stress neighbourhood effects in the early stages of Italian reading development. *The Quarterly Journal of Experimental Psychology*, *66*, 2073-2084.

Thornton, A. M., Iacobini, C., & Burani, C. (1997). *BDVDB: Una base di dati sul vocabolario di base della lingua italiana [BDVDB: A database for the Italian basic dictionary]*. Rome, Italy: Bulzoni.

Footnotes

1. Capital letters indicate the stressed syllable.

2. The exception concerns ultimate stress words, which are marked with a stress diacritic on the last syllable.

3. The percentages reported come from an analysis of BDVDB lemmas (Thornton et al., 1997) where monosyllables were removed

4. Types are word forms and tokens are occurrences of word forms. For example, in the previous sentence there are two tokens each of the types “are”, “word”, and “forms”.

5. There is also a very small number of preantepenultimate stress word forms, which were omitted from our analysis.

6. This may also be due to the fact that in longer words, *-ile* is often preceded by *-b-*. The resulting four-letter ending *–bile*, a derivational suffix analogous to English *-able*, is strongly associated with antepenultimate stress.

7. This finding is qualified by the fact that in texts written for children longer words are prevalently enclitic or third person plural verb forms (e.g., abbandoNAtemi ‘leave me here’; dimentiCArono ‘they forgot’), for which antepenultimate stress is typical.

Supplementary materials

Q2Stress is freely available at http://www.istc.cnr.it/grouppage/databases and comes with functions (see the "scripts" folder) and summary tables (see the "summary tables" folder).

Functions

These functions, written in R, allow the user to query the databases using multiple criteria. R is powerful and open-source and runs on any operative system (To download R: https://www.r-project.org/). Together with four functions, two README files are available, one for R novices (README_ForRNovices.R) and one for R users (README_ForRUsers.R). The functions can be easily run by following the instructions given in the README files.

To investigate word beginnings, the functions called prepbeg_nOfLet.R and beg.R (or CHILD_prepbeg_nOfLet.R and CHILD_beg.R for the child database) should be run. The former creates a database used by the latter for queries. In its basic use, the beg.R script allows the user to specify a word beginning (e.g., *to*-) and a stress pattern (e.g., penultimate stress), and obtain all words starting with that beginning and bearing the specified stress pattern (e.g., *torrone* 'nougat', *tovaglia* 'tablecloth', etc.). However, the variables we took into account are interesting not only *per se* (i.e., how stress patterns distribute according to each word beginning) but also because of the way they interact with each other (i.e., how stress patterns distribute according to grammatical category, number of syllables, and the word ending's CV-structure). Therefore, beg.R allows more detailed queries taking into account other parameters such as, e.g., the number of syllables (for further available parameters see the README files): In doing so, the researcher can run clearer and more refined investigations. Note that the output of each query is saved in a txt file.

The investigation of word ending can be performed by means of two other functions, that is prepFinSeq.R and finSeq.R (CHILD_finSeq.R for the child database), with the former creating the database used by the latter. Then, finSeq.R works as beg.R: By specifying a word ending (e.g., *-ola*)

and a stress pattern (e.g., antepenultimate stress), the functions return all the words ending with that final sequence and bearing that stress pattern (e.g.,*fragola* 'strawberry', *pentola* 'pot', etc.). The research output is saved in a txt file. Also in this case, further parameters (i.e., number of syllables and grammatical category) can be specified in order to refine the research.

Summary tables

Summary tables, provided as tab-delimited text files, report overall counts for word beginnings, word endings, CV-structures of endings, and whole-word CV-structures. Adult and child data are reported in separate folders, each with three folders for word beginnings, word endings, and CV-structures. Type and token data are reported in separate files. Each file reports the percentage and the number of ultimate (ULT), penultimate (PULT), antepenultimate (APULT), and/or preantepenultimate (PAPULT) stress words in which each unit (beginning, ending, CV-structure of ending, or whole-word CV-structure) appears.

The “beginnings” folder includes summary tables for beginnings of one up to four letters. Summary tables are available with both overall counts and separate counts for words with different number of syllables. The “endings” folder includes summary tables for endings and CV-structures of endings in separate subfolders (i.e., “endings” and “CV-endings”, respectively). In the “CV-endings” subfolder, summary tables report stress occurrences for every possible CV-structure of three-letter endings (e.g., VCV, VVV, VVC, CCC, etc.). Summary tables are available with both overall counts and separate counts for each number of syllables as well as each grammatical category. The “endings” subfolder includes summary tables for ending units of one up to four letters. Only three-letter endings with the most frequent CV-structures (i.e., VCV, CCV, VVV, VCC) are reported, with one summary table for each CV-structure. Again, summary tables are available with overall counts as well as with separate counts for each number of syllables and (for three-letter endings only) grammatical category. Finally, the “CV-structures” folder includes summary tables for CV-structures of whole words (with the exception of words with a stress diacritic on the last syllable, which were deemed of little interest for the present purposes). Both overall counts and separate counts for each grammatical category and

number of syllables are available. However, because whole-word CV-structures tend to have the same number of syllables, data for two-syllable words, three-syllable words, etc. are reported in separate files (included in the “types_syllables” and “tokens_syllables” subfolders) rather than a single one.

Table 1. Overall distribution of stress patterns in Italian – data from the adult frequency count.

	Types		Tokens	
Stress pattern	Number	Percentage	Number	Percentage
ultimate	4634	4.93	72707	3.34
penultimate	72167	76.74	1897425	87.05
antepenultimate	17004	18.08	207948	9.54

Note: For each stress pattern it is reported how many words bear it (Number) and their percentage on total words (Percentage), in terms of both types and tokens.

Figure 1. Type and Token distribution of stress patterns in words with different number of syllables (data from the adult frequency count). Occurrences for polysyllabic words with 7 or more syllables have been summed up.

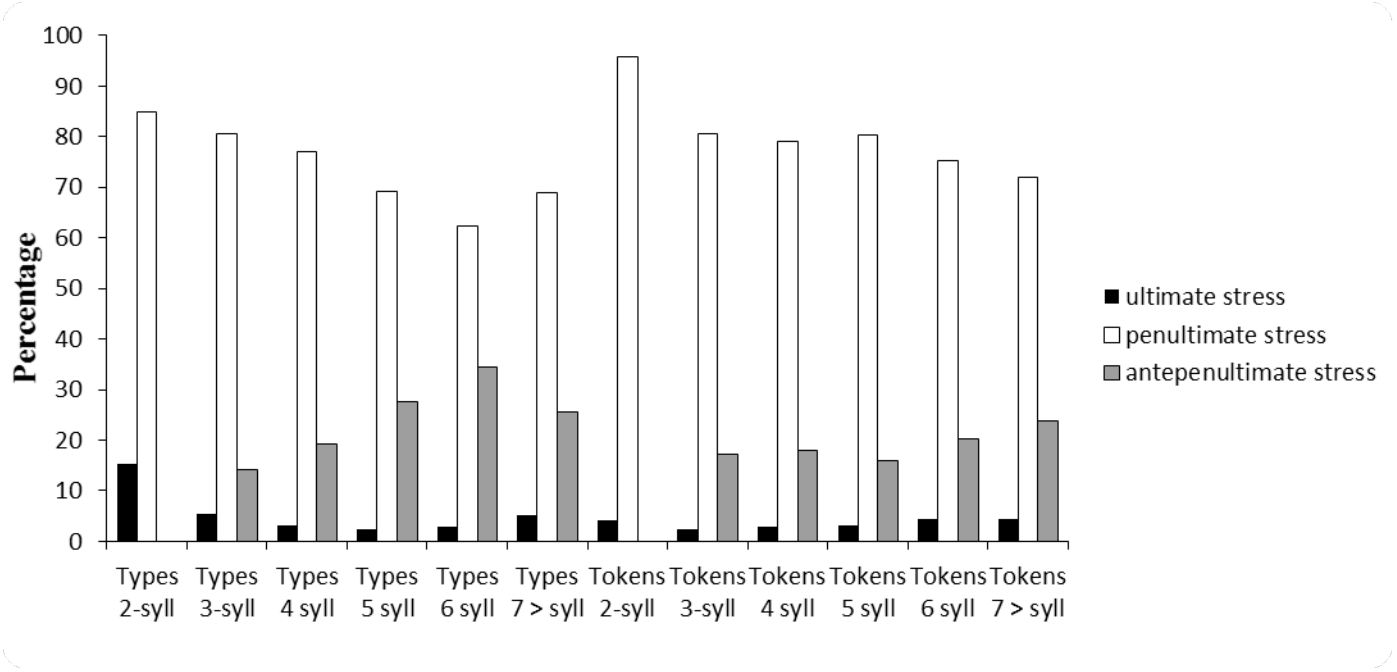

Figure 2. Distribution of stress patterns in words belonging to different grammatical categories (data from the adult frequency count). For each stress pattern it is reported the percentage of words belonging to a given grammatical category that bear it, in terms of both types and tokens. Data for articles, prepositions, pronouns, conjunctions, and other minor grammatical categories, are not reported.

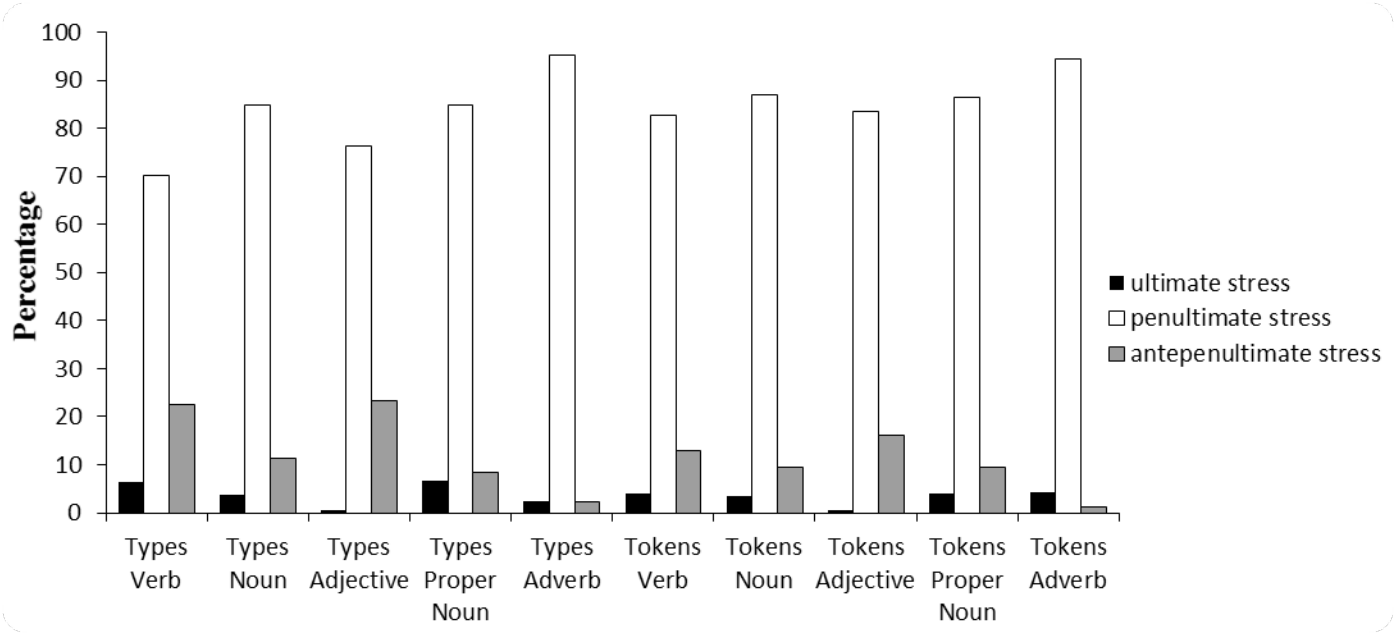

Figure 3. Distribution of stress patterns within words with different ending CV-structures (data from the adult part). For each stress pattern it is reported the percentage of words ending in a given CV-structure that bear it, in terms of both types and tokens. Minor ending CV-structures are not reported.

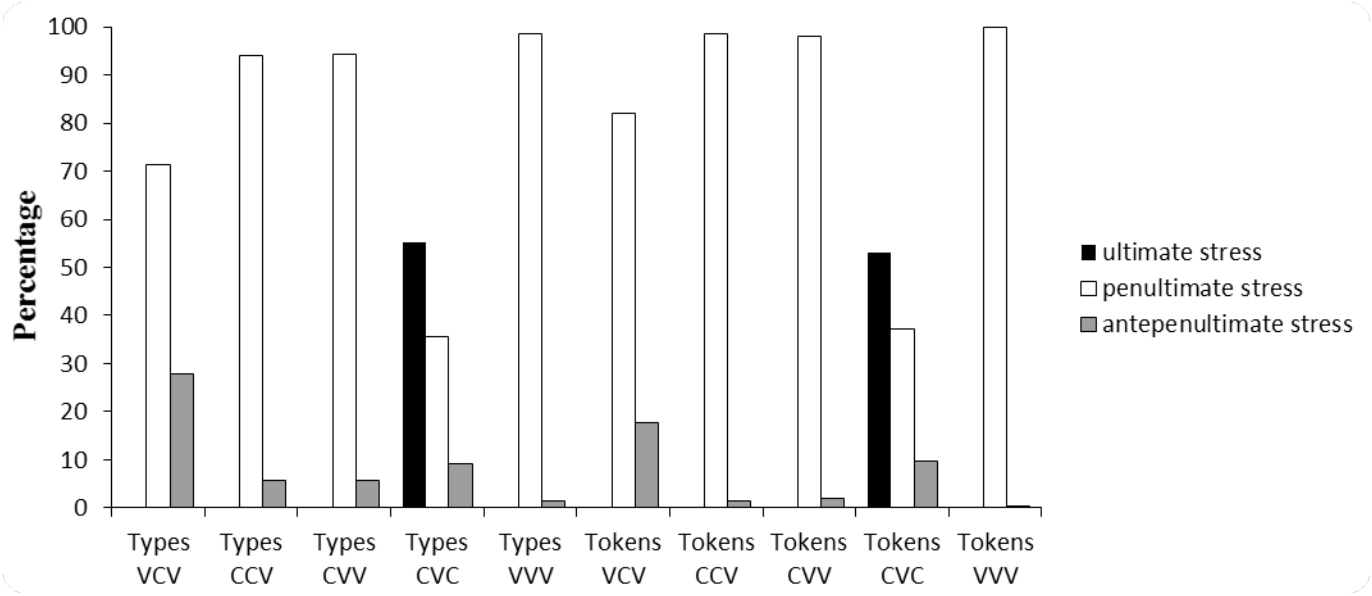

Table 2. Overall distribution of stress patterns in Italian – data from the child-directed frequency count.

	Types		Tokens	
Stress pattern	Number	Percentage	Number	Percentage
ultimate	1316	5.77	12734	4.65
penultimate	16839	73.79	233369	85.25
antepenultimate	4550	19.94	27395	10.01

Note: For each stress pattern it is reported how many words bear it (Number) and their percentage on total words (Percentage), in terms of both types and tokens.

Figure 4. Type and Token distribution of stress patterns within words with different number of syllables (data from the child-directed frequency count). Occurrences for polysyllabic words with 7 or more syllables have been summed up.

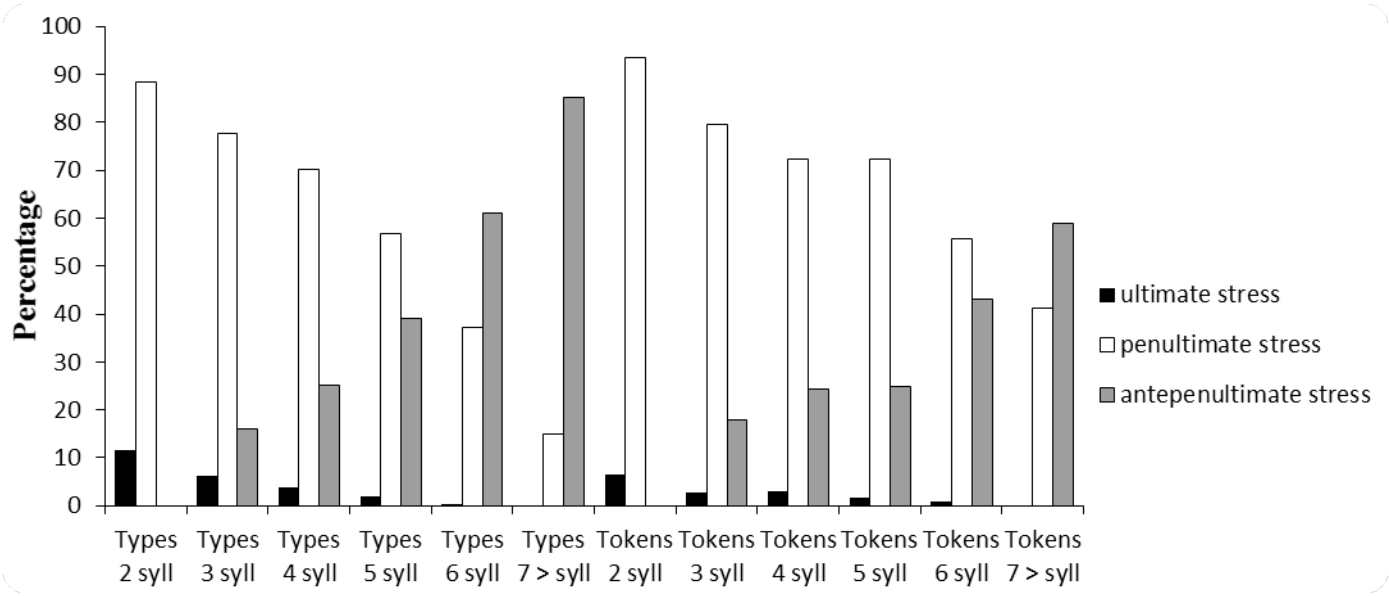

Figure 5. Type and Token distribution of stress patterns within words belonging to different grammatical categories (data from the child-directed frequency count). Proper nouns are absent from this database. Data for articles, prepositions, pronouns, conjunctions, and other minor grammatical categories, are not reported.

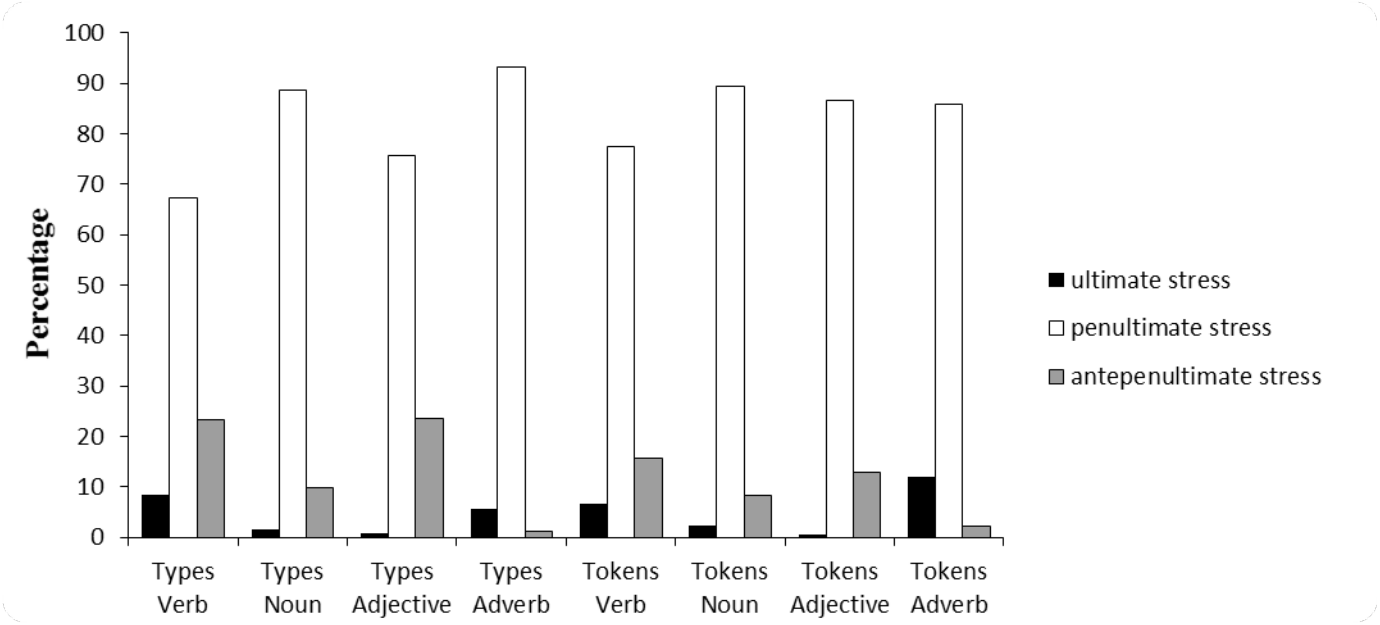

Figure 6. Type and Token distribution of stress patterns within words with different ending CV-structures (data from the child-directed frequency count). Minor ending CV-structures are not reported.

