Behavior Research Methods (2020) 52:2480–2488
https://doi.org/10.3758/s13428-020-01396-2

Lexique-Infra: grapheme-phoneme, phoneme-grapheme regularity, consistency, and other sublexical statistics for 137,717 polysyllabic French words

Manuel Gimenes[1] · Cyril Perret[1,2] · Boris New[3]

Published online: 21 May 2020

Abstract
Psycholinguistic research has shown that both the regularity and consistency of the grapheme-phoneme and phoneme-grapheme correspondences impact word processing. Lexique-Infra is a new database providing infra-lexical statistics for 137,717 French words. The frequencies of the grapheme-phoneme and phoneme-grapheme correspondences as well as other indicators (consistency, regularity, letter frequencies, bigrams, trigrams, phonemes, biphones, and syllables, etc.) are proposed and have been computed from the corpus of subtitles in Lexique 3.83. The aim of this new database is to propose numerous infra-lexical variables based on adult frequencies for a large number of words.

Keywords Phonology · Spelling · Grapheme · Phoneme · Regularity · Consistency · French · Grapheme-phoneme correspondence rules · Phoneme-grapheme correspondence rules.

Introduction

In modern societies, the ability to read and write – literacy – is a fundamental skill for individual and professional development (Nussbaum, 2011; Sen, 1999). Undoubtedly, the psychological mechanisms that support these skills deserve to be further understood. Among a large field of questions, the relationship between orthographic and phonological codes is vital in alphabetic language because it is widely accepted that the processing of orthographic information is influenced by phonological codes (see, for instance, Rastle & Brysbaert, 2006 in English; Ferrand & Grainger, 1993; Roux, McKeeff, Grosjacques, Afonso, & Kandel, 2013, in French). The aim of the present article is to provide a new database for French words. With this new database, we propose a set of infralexical statistics relating to the association between orthographic and phonological infra-lexical codes.

✉ Manuel Gimenes
manuel.gimenes@univ-poitiers.fr

[1] Université de Poitiers, CeRCA/MSHS, Bâtiment A5, 5, rue Théodore Lefebvre, TSA 21103, 86073 Poitiers Cedex 9, France

[2] Université de Poitiers, LMA, Site du Futuroscope - Téléport 2, 11 Boulevard Marie et Pierre Curie, Bâtiment H3 - TSA 61125, F-86073 Poitiers Cedex 9, France

[3] University Grenoble Alpes, University Savoie Mont Blanc, CNRS, LPNC, F-38000 Grenoble, France

Regularity and consistency effects

The relationship between phonological and orthographic information is described by the reciprocal association between an infra-lexical phonological code (i.e., a phoneme: the smallest unit of sound in speech) and an infra-lexical orthographic code (i.e., a grapheme: one or more letters that represent a phoneme in a writing system). This association can be univocal for transparent languages such as Italian or German. In this case, a single phoneme is reciprocally associated with a single grapheme. On the other hand, the French and English orthographies (or writing systems) are fairly opaque. For instance, the phoneme /o/ can be spelled with a large set of graphemes like O, AU, or EAU in French. Moreover, this multiplicity of associations is not necessarily the same between the two types of codes. While the phoneme /o/ can be spelled with a large number of graphemes, the grapheme O is always orally produced with the phoneme /o/. All combinations of these two kinds of associations exist among a variety of alphabetic languages.

Two main indicators are widely used to operationalize these high varieties of associations, namely regularity and consistency. Regularity is a dichotomic variable: a word is regular or irregular. A regular word is one in which all

grapheme-phoneme or phoneme-grapheme correspondences are the most frequent ones in a given language. If at least one correspondence is not the most frequent one, then the word is irregular (Cortese & Simpson, 2000; Protopapas & Vlahou, 2009; Zevin & Seidenberg, 2006). For example, the word “pint” is considered as irregular because the phoneme that corresponds to the grapheme *i* is not the most frequent one. On the contrary, the word “punt” is regular because all the grapheme-phoneme correspondences are the most frequent ones. The second indicator, consistency, refers to the ambiguity between phonological and orthographic codes. Whilst Glushko (1979) defined consistency as a binary variable, most researchers considered it to be a continuous one. Consistency can be seen as statistical information regarding the degree to which a phonological code is related to an orthographic code. For instance, a grapheme-phoneme correspondence with a consistency value of 1 indicates that the given grapheme always corresponds to the given phoneme (the grapheme is always pronounced in the same way). A grapheme-phoneme correspondence with a consistency value of 0.5 means that the given grapheme corresponds to the given phoneme in 50% of the cases (and in the other half of the cases, the given grapheme corresponds to other phonemes). Consistency was often studied in relation to monosyllabic word rhymes (Jared, 2002; Zevin & Seidenberg, 2006). However, consistency can be measured for other units, such as phonemes, graphemes or syllables. In the present article, the indicators are calculated at the grapheme and phoneme levels.

The regularity and consistency effects have been largely studied in visual word processing. Experimental studies have identified that in lexical decision or naming tasks performance is improved for regular words compared to irregular words (Borgwaldt, Hellwig, & de Groot, 2005; Parkin, 1982 ; Parkin & Underwood, 1983 ; Seidenberg, Waters, Barnes, & Tanenhaus, 1984 Stanovich & Bauer, 1978). This result has been observed in various languages, in particular in French (Ziegler, Perry, & Coltheart, 2003), and it has been shown to interact with frequency: regular words are faster to process than irregular words, but only for low-frequency ones (Andrews, 1982; Hino & Lupker, 2000; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Beyond regularity, naming performance in adults is also affected by grapheme-phoneme consistency (Content, 1991; Content & Peereman, 1992; Jared, 1997; Jared, 2002; Peereman, 1995; Seidenberg et al., 1984; Ziegler et al., 2003). For example, Jared (2002) revealed that naming latencies were longer for inconsistent words compared to consistent words. Jared (2002) also manipulated both regularity and consistency in order to know what best characterizes word naming performance. The results showed a clear effect of consistency and a small but significant effect of regularity. Finally, reading performance in adults is also affected by phoneme-grapheme consistency (Grainger & Ziegler, 2008; McKague, Davis, Pratt, & Johnston, 2008; Ziegler, Jacobs, & Stone, 1996). This feedback consistency effect (Stone, Vanhoy, & Van Orden, 1997) was found in a visual lexical decision task: latencies were longer for phoneme-grapheme inconsistent words than for phoneme-grapheme consistent words (all words being consistent in the grapheme-phoneme correspondence).

In handwritten word production, the presence of irregularities made it possible to work on the way in which different processing levels are articulated (Kandel & Perret, 2015; Olive, 2014; Perret & Olive, 2019; Roux et al., 2013). The presence of an irregularity/inconsistency influences access to the spelling codes of a word. For example, when an individual has to write the word “femme” (woman), the first letter "e" is pronounced "a" in French. This situation is at the origin of a conflict between the information offered through the lexical and sublexical pathways. This results in a slowing down of the time course of access to the spelling codes. This proposition has been supported by the results of Bonin et al. (2001). These authors showed that an inconsistency in the initial position of a word to be produced (e.g., oeuf/eeg) increases initialization latencies relative to a consistent/regular word (e.g., ours/bear). However, this influence disappears when the inconsistency/irregularity is in the final position of the word to be produced. Roux et al. (2013, see also Afonso, Alvarez & Kandel, 2015) reported that handwriting durations were longer when the irregularity was in the word final position. More precisely, these authors observed that handwriting characteristics (duration, peak velocity, etc.) were modified for the part of the graphic trace immediately preceding the irregularity. For instance, the writing characteristics of the letter "f" are significantly different in the words "femme" (women) and "fable" (fable). Taken together, these results suggest that, in adults, the difficulty generated by the presence of an irregularity (i.e., conflict between lexical and sublexical information) would be managed either before the start of handwritten production if the irregularity is in the initial position or during writing if the irregularity is in the final position. Then the processes dedicated to access to orthographic codes (i.e., central processes) and those dedicated to the preparation and realization of the motor trace (i.e., peripheral processes) are carried out in parallel (Kandel & Perret, 2015 ; Olive, 2014 ; Roux et al., 2013).

Regularity and consistency in theoretical models

Several theoretical models have been proposed to explain how certain variables (such as frequency, regularity, and consistency) affect performance in visual word recognition. These models can be classified into two main types. The first is represented by the dual-route cascaded (DRC) model (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001); the second is represented

by the parallel distributed processing (PDP) model (Plaut, 1996; Seidenberg & McClelland, 1989).

The DRC model proposes two routes: a lexical route and a sublexical route. The lexical route is based on an interactive activation procedure (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) that activates a phonological code in the mental lexicon from a visual word code. In contrast, the sublexical route is based on grapheme-phoneme correspondence rules: each grapheme is converted into a corresponding phoneme by the application of a set of rules. During word recognition, the two routes work simultaneously and convey information to a common phoneme system. The DRC model can interpret both regularity and consistency effects, but with different procedures for each. The regularity effect is the result of a conflict between the lexical and sublexical routes. When an irregular word is read, the lexical route produces the correct pronunciation while the sublexical route does not. In this case, two phonemes are activated and inhibit each other. This conflict increases reaction time, compared to a regular word in which both routes produce the correct pronunciation. The consistency effect is interpreted differently in the DRC model and is supposed to arise within the lexical route. In this lexical route, information at different levels (e.g., letter, grapheme, word) is activated in cascade. For example, the sequence of letters "ood" activate words that contain them (e.g., mood, good, wood). When words with different pronunciations are activated, it results in competition in the phoneme system and increases reaction times. The DRC model can also interpret the frequency–regularity interaction observed in some studies (Andrews, 1982; Hino & Lupker, 2000), the fact that the regularity effect is in general observed only for low-frequency words and not for high-frequency words. The interpretation is in terms of time course of processing along the two routes. For high-frequency words, the phonological information is quickly retrieved from the lexical route. When responding, the information from the sublexical route is not yet available and there is no conflict and so no regularity effect. On the contrary, for low-frequency words, information from both routes is available at the same time, resulting in a conflict and so in a regularity effect.

The PDP model is based on a connectionist framework. In the model proposed by Plaut, McClelland, Seidenberg and Patterson (1996), there is an orthographic input level, a phonological output level and a semantic level representing the frequency of the orthographic input. In the implemented version of the model, each orthographic unit is connected to phonological ones via hidden units. Each connection is linked to a weighted value that is modulated during a learning phase. The grapheme-phoneme knowledge is therefore represented as weights on connections between orthographic units and phonological units. There is not a unique representation for each word: the same set of units and connections is used for all words. Additionally, there are no grapheme-phoneme correspondence rules. Because of this architecture, naming a given word is influenced by the pronunciation of orthographically similar words. The PDP model then predicts consistency effects: letter patterns that are always pronounced the same way in a high number of words will be read aloud faster than letter patterns that are pronounced differently in different words. According to this approach, regularity effects are actually due to grapheme-phoneme inconsistencies. The PDP model can also interpret the interaction between frequency and regularity: word frequency can override consistency or regularity effects (the frequency of a given word affects all of its connections whereas inconsistency affects only a portion of these connections). Consequently, a regularity effect is predicted only for low-frequency words.

Statistical databases

Because of the regularity and consistency effects observed in scientific research, it seems important to have infralexical statistics about grapheme-phoneme and phoneme-grapheme correspondences. In English, numerous databases are available (Berndt, D'Autrechy, & Reggia, 1994; Berndt, Reggia, & Mitchum, 1987; Gontijo, Gontijo, & Shillcock, 2003; Hanna, Hanna, Hodges, & Rudorf, 1966; Wijk, 1966; Ziegler, Stone, & Jacobs, 1997). In French, to the best of our knowledge, three grapheme-phoneme and/or phoneme-grapheme consistency databases have been proposed and are still available (Peereman & Content, 1999; Peereman, Lété, & Sprenger-Charolles, 2007; Ziegler et al., 1996).

Ziegler et al. (1996) provided a statistical database indicating the degree of inconsistency in both the grapheme-to-phoneme and phoneme-to-grapheme directions. The statistical analysis of the inconsistency was based on 1843 monosyllabic Brulex words (Content, Mousty, & Radeau, 1990) and was calculated for all orthographic and phonological bodies (in a monosyllabic word, the onset is the initial sequence of consonants; the body corresponds to everything else in the word). Peereman and Content (1999) provided a statistical database describing the relationships between orthography and phonology for 2449 single-syllabic French words (from Brulex). Three categories of variables were proposed: the consistency of grapheme-to-phoneme and phoneme-to-grapheme associations, the frequency of orthographic and phonological correspondences, and lexical neighborhood. For each variable, the authors proposed a "type" frequency and a "token" frequency. The type frequency corresponds to the number of occurrences of a given grapheme-phoneme association in a corpus, while token frequency weighted the number of occurrences by the frequency of the words. Similarly, Peereman et al. (2007) provided a statistical database on several infra-lexical variables (syllable, grapheme-to-phoneme mappings, bigrams)

and lexical variables (lexical neighborhood, homophony and homography). The data were based on lexical frequencies calculated from a corpus of textbooks (Manulex). Statistics were provided for 45,080 words (mono- and poly-syllabic), including both type and token frequencies. The authors calculated the association frequencies and also the consistency indices for grapheme-phoneme and phoneme-grapheme associations. These frequency and consistency indices were calculated according to the position of the units in the word (initial, middle, final).

These three French language databases are a crucial step forward, but they still have some limitations. Ziegler et al. (1996) and Peereman and Content (1999) had only monosyllabic words and calculations were made at the rime level (and not at the grapheme level, for example). Peereman et al. (2007) based their calculations on a corpus for children. Our study aims to advance beyond these limits. More precisely, the main objective of this article is to provide infra-lexical statistics for a very large number of polysyllabic words (more than 130,000) from an adult corpus. For the 137,717 words in Lexique 3.83, we propose the frequency, regularity, and consistency of the overall grapheme-phoneme and phoneme-grapheme correspondences of a word, but also for each grapheme or phoneme by distinguishing its position in the word (beginning, middle, and end), as proposed in Manulex-infra (Peereman et al., 2007). Several new indicators of consistency and regularity are also proposed: the number of irregularities in a word, the position of the irregularity, the average complexity of the graphemes of a word and the frequency of the lowest inconsistency in a word.

Method

Lexique-Infra is based on 137,717 words in Lexique 3.83 (New, Pallier, Brysbaert, & Ferrand, 2004). Lexique 3.83 was created from a corpus of subtitles of 9474 movies and television series consisting of 52 million words and a corpus of 218 books (novels) consisting of 14.7 million words. We have chosen to use the corpus of subtitles (surface frequencies) as a basis for calculating our various subtitle frequency indices because the literature has regularly shown that subtitle frequencies predict reaction times better than book frequencies even in visual word recognition. For example, in French, New, Brysbaert, Veronis & Pallier (2007) demonstrated that frequencies from a corpus of subtitles predicted reaction times better than frequencies from a corpus of books (see Brysbaert & New, 2009 for a similar study in English). All entries in the Lexique database were used, with the exception of compound words and acronyms.

Lexique's phonological codes were used (more details are available here). The phonological codes used include 16 vowels, three glides and 19 consonants. The hash (#) was also used to indicate the graphemes that were not pronounced. For example, in the French word "corps" ("body" in English), the grapheme "ps" is paired with the phoneme "#".

Graphemic segmentation

As the segmentation of each word into graphemes is not proposed by Lexique, we had to calculate it. To do this, we have tried, as far as possible, to match a given phoneme to a given grapheme (the list of these 348 matching rules is provided in the file "Lexique.Infra.Rules.xlsx"). The algorithm scanned the word from the first letter to the last. For each letter, it began by looking for the longest grapheme (in number of letters) associated with the corresponding phoneme. If it did not find a complex grapheme, the search focused on simpler graphemes and so on until it identified the right grapheme. For 197 words with exceptional pronunciation (e.g., words of foreign origin), the segmentation was performed manually.

Several studies have found that morphological knowledge has an impact on learning to read and write (Casalis & Louis-Alexandre, 2000; Clin, Wade-Woolley, & Heggie, 2009; Colé, Bouton, Leuwers, Casalis, & Sprenger-Charolles, 2012; Deacon & Kirby, 2004; Mahony, Singson, & Mann, 2000; Rastle, Davis, & New, 2004; Shankweiler et al., 1995). With the flexional morphology of conjugated forms being relatively complex in French, the segmentation applied to the endings of conjugated verbs differed from that applied to other grammatical categories. An example is the following: the words "jouait" (conjugated verb, "played" in English) and "bienfait" (common word, "benefaction" in English) both contain the grapheme "ait", which is pronounced /ɛ/. Our algorithm will segment "jouait" as "j.ou.ait" because the morpheme "ait" indicates the progressive preterit of the third person in French, whilst "bienfait " will be segmented as " b.i.en.f.ai.t " because the silent "t" indicates morphological information ("bienfaiteur" is a common noun derived from "bienfait"). Another example is the following: the words "ballons" (common noun, "balloons" in English) and "pouvons" (conjugated verb, "can" as in "we can" in English) both contain the grapheme "ons", which is pronounced /˜/. Our algorithm will segment "ballons" as "b.a.ll.on.s" because the morpheme "s" indicates the plural in French, whilst "pouvons" will be segmented as "p.ou.v.ons" because the morpheme "ons" indicates the first person of the plural in French conjugation of the present.

Infralexical unit frequencies

With regard to orthographic forms, we calculated the frequencies of letters, bigrams (a sequence of two letters), trigrams (a sequence of three letters) and graphemes according to their position in the word (initial, middle, final) by type and by token. For phonological forms, we calculated the

frequencies of phonemes, biphones (a sequence of two phonemes) and syllables according to their position in the word (initial, middle, final). Such frequencies have been calculated by type and token and have also been averaged for each word ("Lexique.Infra.Freq.Let.Bigr.Trig.Syl.Phon.Biph_1.1.xlsb). A statistical description of these infralexical unit frequencies can be found in Table 1. The main objective of the article being the problem with the link between orthographic and phonological codes, we will not more comprehensively discuss descriptors that are either only orthographic or only phonological.

Description of Lexique-Infra variables

Once the segmentation was completed, several indicators were calculated. Some have already been proposed in the literature, whereas others are novel findings. First, we calculated the frequency of each existing grapheme-phoneme association in our corpus ("Lexique.Infra.Corresp.Grapheme-Phoneme_1.11.xlsb): this is simply the number of occurrences of a given grapheme-phoneme association. In the second step, we calculated the grapheme-phoneme consistency by calculating the ratio between the grapheme-phoneme association frequency of the pair in question and the grapheme frequency. We subsequently calculated the phoneme-grapheme consistency by calculating the ratio between the frequency of the phoneme-grapheme association of the pair under consideration and the frequency of the phoneme. These indices were calculated by type and token and according to the initial, middle, and final positions. We decided to use the distinction between the initial, middle and final positions following the work of Peereman et al. (2007) in Manulex-infra. Due to the derivational morphology of French, word endings are often silent, so spelling is less transparent. To better characterize the grapheme-phoneme correspondences of French, frequency and consistency were computed as a function of their position (initial, middle, final) in the words.

For example, suppose that in the initial position, the grapheme "ch" appears 15 times in our corpus, the phoneme /S/ appears 20 times, the pair ch-/S/ (as in “chocolat” [“chocolate”]) appears ten times in our corpus and the pair ch-/k/ (as in “chaos” [“chaos”]) appears five times. For the pair "ch-/S/", its grapheme-phoneme consistency will be 10/15 = 67%, while its phoneme-grapheme consistency will be 10/20 = 50%.

For each word, we calculated the grapheme-phoneme and phoneme-grapheme consistencies of each of its graphemes (for instance: château = ch-/S/.â-/a/.t-/t/.eau-/o/ = “0.99 1.00 0.95 1.00” for its grapheme-phoneme consistency by token). We then calculated the average complexity of the graphemes, i.e., the average number of letters composing each grapheme. We also indicated the minimum consistency of the word (in this example, 0.95). Finally, we indicated the number of irregular graphemes for which the association was not the most frequent and the precise position of the first irregular grapheme. These indices were also calculated by type and token and according to the initial, middle, and final positions. In order to have reference points on the variables presented in Lexique-Infra, users can refer to Table 2 for a statistical description of each variable.

Finally, we wanted to compare our indicators with those given by Manulex. To do this, we calculated the correlations between our consistency indicators and those of Manulex. We therefore matched all Manulex words having the same phonological representation as Lexique words, which allowed us to obtain the consistency data of 36,469 words. In general, the

Table 1 Descriptive statistics for the main infralexical unit frequencies

		Mean	Median	SD	Min	Max
Letter frequency	Type	51,763.16	52,773.2	9614.00	1	78,738.2
	Token	122,166.67	124,341.8	20,585.90	33.1	182,585.9
Bigram frequency	Type	5618.04	5587.2	2320.64	1	14,571.7
	Token	11,058.64	10,724.4	5072.88	0	40,640.40
Trigram frequency	Type	920.95	745.4	723.58	1	5744.00
	Token	1427.29	1115.7	1201.84	0	22,895.50
Phoneme frequency	Type	23,984.19	24,537.1	6894.87	10	46,560.20
	Token	45,276.77	45,533.45	10,483.57	797.5	88,344.4
Biphone frequency	Type	2000.77	1980.8	982.41	1	7738
	Token	3085.98	2904.4	1727.42	0	29,871.2
Grapheme frequency	Type	25,995.78	26,754.8	8183.21	1	51,279.8
	Token	55,746.73	55,226.7	18,557.31	3.8	155,707
Syllable frequency	Type	1259.32	980.7	1098.58	1	8630
	Token	3587.63	1653.85	5167.49	0	50,296

Table 2 Descriptive statistics for the main Lexique-Infra variables

			Mean	Median	SD	Min	Max
Grapheme-phoneme consistency	Mean frequency	Type	0.88	0.89	0.09	0.03	1.00
		Token	0.84	0.85	0.09	0.19	1.00
	Initial frequency	Type	0.95	1.00	0.14	0.00	1.00
		Token	0.93	1.00	0.18	0.00	1.00
	Middle frequency	Type	0.83	0.85	0.13	0.00	1.00
		Token	0.82	0.83	0.13	0.00	1.00
	Final frequency	Type	0.99	1.00	0.09	0.00	1.00
		Token	0.87	1.00	0.20	0.00	1.00
	Freq mini	Type	0.51	0.51	0.28	0.00	1.00
		Token	0.46	0.42	0.25	0.00	1.00
Grapheme-phoneme regularity	Nb irregularities	Type	0.69	0.00	0.88	0.00	8.00
		Token	1.03	1.00	1.01	0.00	9.00
	Position first irregularity	Type	1.87	0.00	2.39	0.00	15.00
		Token	2.41	2.00	2.30	0.00	14.00
Phoneme-grapheme consistency	Mean frequency	Type	0.73	0.75	0.11	0.00	1.00
		Token	0.69	0.70	0.12	0.00	1.00
	Initial frequency	Type	0.90	0.99	0.21	0.00	1.00
		Token	0.82	0.99	0.27	0.00	1.00
	Middle frequency	Type	0.78	0.80	0.16	0.00	1.00
		Token	0.74	0.76	0.15	0.00	1.00
	Final frequency	Type	0.35	0.32	0.24	0.00	1.00
		Token	0.29	0.30	0.25	0.00	1.00
	Freq mini	Type	0.21	0.18	0.18	0.00	1.00
		Token	0.17	0.10	0.17	0.00	1.00
Phoneme-grapheme regularity	Nb irregularities	Type	1.30	1.00	0.96	0.00	8.00
		Token	1.43	1.00	0.96	0.00	8.00
	Position first irregularity	Type	3.62	3.00	2.82	0.00	17.00
		Token	3.75	3.00	2.79	0.00	16.00
Mean grapheme complexity			1.27	1.25	0.22	1.00	5.00

correlations are high (all r >.70). The dot plots (see Fig. 1) clearly demonstrate that although the two databases have some differences, they also have a strong overlap. These differences can be explained by the fact that the corpora from which these consistencies have been established are quite different (1.9 million words based on 54 manuals for children for Manulex-Infra vs. 52 million words based on 9474 movies and television series for adults for Lexique-Infra).

Discussion

It is important to discover the infra-lexical characteristics of words in French since many studies have revealed, for example, that grapheme-phoneme consistency influences reading performance (Content, 1991; Content & Peereman, 1992; Jared, 1997; Jared, 2002; Peereman, 1995; Seidenberg et al., 1984; Ziegler et al., 2003). The objective of this paper is to make available, for all the words in Lexique 3.83, several infra-lexical statistical indicators. First, for each word, we give the grapheme-phoneme and phoneme-grapheme consistencies averaged over the whole word, as well as for each grapheme and each phoneme, taking into account its position in the word. We also indicate the number of grapheme-phoneme and phoneme-grapheme irregularities and their position in the word. Finally, new indices are proposed such as the average complexity of the graphemes in the word or the minimum frequency of the word's grapheme-phoneme or phoneme-grapheme associations. All of these statistics are available by token and by type.

A critical point of Lexique-Infra is that statistics are proposed for polysyllabic words, which was not the case in

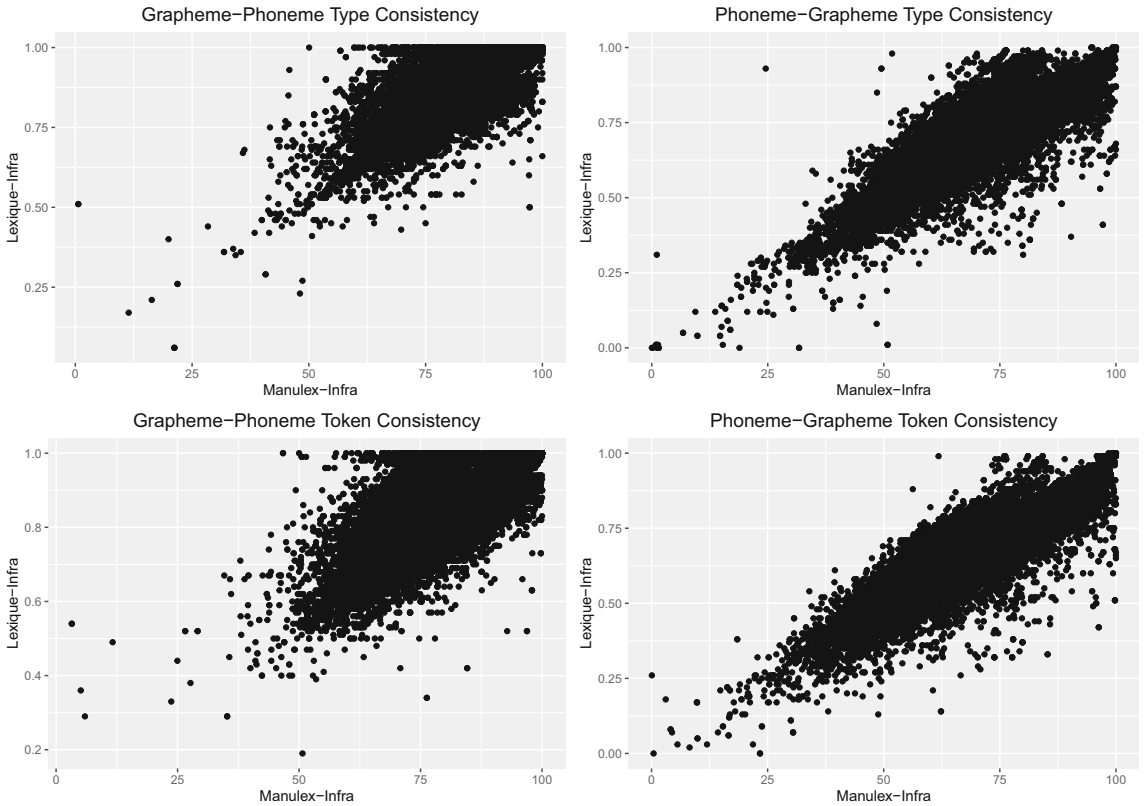

Fig. 1 Dot plots comparing Lexique-Infra and Manulex-Infra consistencies

previous French studies (Ziegler et al., 1996; Peereman and Content, 1999). Only Peereman et al. (2007) provided data for words with several syllables, but the corpus used was textbooks. A major interest of this article is to propose statistics based on a much more representative corpus of adult language. Another difference is that Manulex-infra is offered for 45,080 words while Lexique-Infra is offered for more than 137,000 words.

Lexique-Infra will be useful for better controlling regularity and consistency effects in experimental studies in psycholinguistics, particularly when adult participants are tested. Our database could also be used when researchers wish to manipulate regularity and consistency effects. For instance, the interaction between regularity and frequency has commonly been studied: in general, the regularity effect is weak or null for high-frequency words but greater for low-frequency ones (Jared, 1997; Seidenberg, 1985; Taraban & McClelland 1987). This type of effect could be tested in French with Lexique-Infra. Another example of using our database could be computing regression analyses to predict behavioral performance in naming or lexical decision tasks in megastudies: the regularity and consistency variables in Lexique-Infra could be entered as predictors in statistical models. Finally, the database can be used to create material to accurately evaluate reading and writing difficulties in adults (by selecting, for instance, low-consistency and high-consistency words).

Having individually calculated both the grapheme-phoneme and phoneme-grapheme directions will allow Lexique-Infra to be used in comprehension and production studies. As far as written production is concerned, we briefly described in the Introduction how the manipulation of the presence or absence of an irregularity in a word facilitated interest in the parallel treatment of the central and peripheral levels (Olive, 2014; Perret & Olive, 2019). However, the work in French was based either solely on monosyllabic words (Bonin et al., 2001) or on irregularity calculations from databases for children (Roux et al., 2013). The values provided by our database should improve this methodological aspect. In addition, having a more precise tool with consistency should also allow us to construct experiments to improve our understanding of the interaction between the central and peripheral levels. In particular, this tool should make it possible to explore the moment of conflict management between lexical and sublexical information according to the position of the inconsistency in polysyllabic words, as well as according to the degree of inconsistency.

As stated in the Method section, the graphemic segmentation was not always the same depending on morphological information. For example, the sequence letters "ons" was considered as one grapheme at the end of a verb but as two graphemes at the end of a noun. Two words that have the same pronunciation can therefore have different graphemic segmentations. A question is then raised: in word recognition tasks (such as lexical decision tasks or naming tasks) in which participants are presented with a single word without any context, how would differing consistencies affect response latencies? This question is outside the scope of the present study. However, the Lexique-Infra material could be used to experimentally test this question.

Acknowledgements We thank the authors of Manulex-infra for their work on sublexical frequencies in children. Indeed, the article and the database were an important source of both motivation and inspiration for the realization of the present article. This work was supported by the French National Research Agency (Grant ANR ECRITURE 14-C30-0013-01, awarded to Cyril Perret).

Open Practices Statement The database is downloadable and will be searchable online on http://lexique-infra.lexique.org. This will allow, for example, a user to select all six-letter words beginning with the letter "a" with an irregularity. It will also be possible, for example, to request all words with an average consistency of 90% in order to have words with a good overall average consistency. Two examples are subsequently provided, but the possibilities of queries by matching Lexique 3 and Lexique-Infra or Lexique-Infra and other databases are endless. The scripts (made in Perl) having been used to generate the database are also available.

None of the experiments were preregistered.

References

Afonso, O., Álvarez, C. J., & Kandel, S. (2015). Effects of grapheme-to-phoneme probability on writing durations. *Memory & Cognition, 43(4),* 579–592

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory & Cognition*, *10*(6), 565–575.

Berndt, R. S., D'Autrechy, C. L., & Reggia, J. A. (1994). Functional pronunciation units in English words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 977–991.

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers*, *19*(1), 1–9.

Bonin, P., Peereman, R., & Fayol, M. (2001). Do phonological codes constrain the selection of orthographic codes in written picture naming? *Journal of Memory and Language*, *45*(4), 688–720.

Borgwaldt, S. R., Hellwig, F. M., & de Groot, A. M. (2005). Onset entropy matters–Letter-to-phoneme mappings in seven languages. *Reading and Writing*, *18*(3), 211–229.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Casalis, S., & Louis-Alexandre, M. F. (2000). Morphological analysis, phonological analysis and learning to read French: A longitudinal study. *Reading and Writing*, *12*(3), 303–335.

Clin, E., Wade-Woolley, L., & Heggie, L. (2009). Prosodic sensitivity and morphological awareness in children's reading. *Journal of Experimental Child Psychology*, *104*(2), 197–213.

Colé, P., Bouton, S., Leuwers, C., Casalis, S., & Sprenger-Charolles, L. (2012). Stem and derivational-suffix processing during reading by French second and third graders. *Applied Psycholinguistics*, *33*(1), 97–120.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589.

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*(1), 204.

Content, A. (1991). The effect of spelling-to-sound regularity on naming in French. *Psychological Research*, *53*(1), 3–12.

Content, A., Mousty, P., & Radeau, M. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'année Psychologique*, *90*(4), 551–566.

Content, A., & Peereman, R. (1992). Single and multiple process models of print to speech conversion. In J. Alegria, D. Holender, J. Junca de Morais, & M. Radeau (Eds.), *Analytic approaches to human cognition* (pp. 213–236). Amsterdam: North-Holland.

Cortese, M. J., & Simpson, G. B. (2000). Regularity effects in word naming: What are they? *Memory & Cognition*, *28*(8), 1269–1276.

Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics*, *25*(2), 223–238.

Ferrand, L., & Grainger, J. (1993). The time course of orthographic and phonological code activation in the early phases of visual word recognition. *Bulletin of the Psychonomic Society*, *31*(2), 119–122.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 674.

Gontijo, P. F., Gontijo, I., & Shillcock, R. (2003). Grapheme-phoneme probabilities in British English. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 136–157.

Grainger, J., & Ziegler, J. (2008). Cross-code consistency effects in visual word recognition. In L. Grigorenko, and A. Naples (Eds.), *Single-Word Reading: Biological and Behavioral Perspectives* (pp. 129–157). Mahwah, NJ: Lawrence Erlbaum Associates.

Hanna, P. R., Hanna, J. S., Hodges, R. E., & Rudorf, E. H. (1966). Phoneme-grapheme correspondences as cues to spelling improvement. Washington, DC: U.S. Department of Health, Education, and Welfare.

Hino, Y., & Lupker, S. J. (2000). Effects of word frequency and spelling–to–sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 166.

Jared, D. (1997). Spelling–sound consistency affects the naming of high-frequency words. *Journal of Memory and Language*, *36*(4), 505–529.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*(4), 723–750.

Kandel, S., & Perret, C. (2015). How does the interaction between spelling and motor processes build up during writing acquisition? *Cognition*, *136*, 325–336.

Mahony, D., Singson, M., & Mann, V. (2000). Reading ability and sensitivity to morphological relations. *Reading and Writing*, *12*(3), 191–218.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375.

McKague, M., Davis, C., Pratt, C., & Johnston, M. B. (2008). The role of feedback from phonology to orthography in orthographic learning:

an extension of item-based accounts. *Journal of Research in Reading*, *31*(1), 55–76.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*(4), 661–677.

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

Nussbaum, M. C. (2011). *Creating capabilities*. Cambridge, MA: Harvard University Press.

Olive, T. (2014). Toward an Incremental and Cascading Model of Writing: A review of research on writing processes coordination. *Journal of Writing Research*, *6*, 173–194.

Parkin, A. J. (1982). Phonological recoding in lexical decision: Effects of spelling-to-sound regularity depend on how regularity is defined. *Memory & Cognition*, *10*(1), 43–53.

Parkin, A. J., & Underwood, G. (1983). Orthographic vs. phonological irregularity in lexical decision. *Memory & Cognition*, *11*(4), 351–355.

Peereman, R. (1995). Naming regular and exception words: Further examination of the effect of phonological dissension among lexical neighbours. *European Journal of Cognitive Psychology*, *7*(3), 307–330.

Peereman, R., & Content, A. (1999). LEXOP: A lexical database providing orthography-phonology statistics for French monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *31*(2), 376–379.

Peereman, R., Lété, B., & Sprenger-Charolles, L. (2007). Manulex-infra: Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods, 39*(3), 579–589.

Perret, C., & Olive, T. (2019). *Spelling and Writing Words: Theoretical and Methodological Advances*. Leiden: Brills Edition.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language*, *52*(1), 25–82.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56.

Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods*, *41*(4), 991–1008.

Rastle, K., & Brysbaert, M. (2006). Masked phonological priming effects in English: Are they real? Do they matter? *Cognitive Psychology*, *53*(2), 97–145.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090–1098.

Roux, S., McKeeff, T. J., Grosjacques, G., Afonso, O., & Kandel, S. (2013). The interaction between central and peripheral processes in handwriting production. *Cognition*, *127*(2), 235–241.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*(1), 1–30.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior*, *23*(3), 383–404.

Sen, A. (1999). Commodities and capabilities. *OUP Catalogue*.

Shankweiler, D., Crain, S., Katz, L., Fowler, A. E., Liberman, A. M., Brady, S. A., … Stuebing, K. K. (1995). Cognitive profiles of reading-disabled children: Comparison of language skills in phonology, morphology, and syntax. *Psychological Science*, *6*(3), 149–156.

Stanovich, K. E., & Bauer, D. W. (1978). Experiments on the spelling-to-sound regularity effect in word recognition. *Memory & Cognition*, *6*(4), 410–415.

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language, 36*(3), 337–359.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language, 26*(6), 608–631.

Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition*, *13*(6), 557–572.

Wijk, A. (1966). Rules of Pronunciation for the English Language: An account of the relationship between English spelling and pronunciation (Vol. 12). London: Oxford University Press.

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language*, *54*(2), 145–160.

Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, & Computers*, *28*(4), 504–515.

Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What is the pronunciation for -ough and the spelling for/u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, & Computers, 29*(4), 600–618.

Ziegler, J. C., Perry, C., & Coltheart, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychonomic Bulletin & Review*, *10*(4), 947–953.