COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13118

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13118

Gepo with a G, or Jepo with a J? Skilled Readers Generate Orthographic Expectations for Novel Spoken Words Even When Spelling is Uncertain

Mina Jevtović,[a,b] Alexia Antzaka,[a] Clara D. Martin[a,c]

[a]*Basque Center on Cognition, Brain and Language (BCBL), Donostia-San Sebastián, Spain*
[b]*University of the Basque Country (UPV/EHU), Leioa*
[c]*Basque Foundation for Science (Ikerbasque), Bilbao*

Received 26 June 2021; received in revised form 8 February 2022; accepted 9 February 2022

Abstract

English-speaking children and adults generate *orthographic skeletons* (i.e., preliminary orthographic representations) solely from aural exposure to novel words. The present study examined whether skilled readers generate orthographic skeletons for all novel words they learn or do so only when the words have a unique possible spelling. To that end, 48 Spanish adults first provided their preferred spellings for all novel words that were to appear in the experiment. Critically, *consistent words* had only one, while *inconsistent words* had two possible spellings. Two weeks later, they were trained on the pronunciations of the novel words through aural instruction. They then saw the spellings of these newly acquired words, along with a set of untrained words, in a self-paced sentence reading task. Participants read previously acquired consistent and inconsistent words presented in their preferred spellings faster than inconsistent words with unpreferred spellings. Importantly, no differences were observed in reading untrained consistent and inconsistent words (either preferred or unpreferred). This suggests that participants had generated orthographic skeletons for trained words with two possible spellings according to their individual spelling preferences. These findings provide further evidence for the orthographic skeleton account and show that initial orthographic representations are generated even when the spelling of a newly acquired word is uncertain.

Keywords: Orthographic representations; Orthographic consistency; Word recognition; Reading; Spelling; Word learning

Correspondence should be sent to Mina Jevtović, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 20009 Donostia, Spain. E-mail: mjevtovic001@ikasle.ehu.eus

1. Introduction

Children's first encounter with words usually occurs in spoken language (aurally). Once they begin to read and acquire phoneme-to-grapheme conversion rules, children become exposed to written language (in orthographic form). Thus, when children first encounter a written word, they are already familiar with its meaning and pronunciation. On some occasions, such as acquiring novel words through spoken language, this can also happen with skilled adult readers. Evidence suggests that even when processing spoken language or, more precisely, single auditory words, orthographic information associated with these words is coactivated (Chéreau, Gaskell, & Dumay, 2007; Perre & Ziegler, 2008). In the same vein, it has been proposed that knowledge of phoneme-to-grapheme mappings can be used to generate preliminary orthographic representations of already familiar spoken words, even before they have been seen in writing (see Stuart & Coltheart, 1988).

Johnston, McKague, and Pratt (2004; see also McKague, Davis, Pratt, & Johnston, 2008) presented the first evidence that *orthographic expectations*[1] can be generated solely from aural and prior to any visual exposure. When we read, we sound out unfamiliar words by mapping letters to sounds (see Share, 1995). The authors assumed that a reverse mechanism allows us to map sounds to letters, this way generating orthographic expectations when we aurally acquire novel words.

In three masked-priming visual lexical decision task experiments, Johnston et al. (2004) showed that English adults had encoded orthographic representations of novel words previously acquired through aural instruction. To tease out any potential effects of orthographic priming during access to novel visual words, they created four different *prime-target* conditions, which overlapped in either phonological form (e.g., <vornce> before <VAUNCE>), orthographic form (e.g., <veenie> before <VAUNCE>), both (e.g., <vaunce> before <VAUNCE>) or neither (e.g., <mellop> before <VAUNCE>). Novel words preceded by identical primes (i.e., those with both phonological and orthographic overlaps) were processed faster than those preceded by phonological only primes, thus demonstrating that representations of novel spoken words are not purely phonological. In addition to replicating priming effects found for familiar words (Forster & Davis, 1984), this showed that representations of novel words were automatically accessed using the same recognition mechanisms employed when accessing already existing orthographic representations of familiar words. Finally, and crucially for the authors' conclusions, there was no difference in processing novel words preceded by phonological primes and those preceded by purely orthographic primes. The latter were, however, processed faster than novel words preceded by orthographically different primes, that is, primes that were spelled using a completely different set of letters. The absence of phonologically mediated priming alongside the significant differences observed between orthographically similar and dissimilar primes led the authors to conclude that these English adults had already generated orthographic representations of novel spoken words. However, it remained unclear whether a single expectation was generated for each word or whether multiple expectations for alternative spelling patterns were considered.

Recent evidence supporting the idea that preliminary orthographic representations are generated as a result of aural instruction led Wegener and colleagues to propose the *orthographic* *skeleton hypothesis* (Wegener et al., 2018; Wegener, Wang, Nation, & Castles, 2020). Noting the positive link between children's oral vocabulary and their future reading skills (Duff & Hulme, 2012; McKague, Pratt, & Johnston, 2001; Nation & Cocksey, 2009; Nation & Snowling, 2004), the authors hypothesized that this link could be mediated by a mechanism similar to the one described by Johnston et al. (2004). They asked whether knowledge of phoneme-to-grapheme mappings could support the creation of orthographic representations prior to any visual encounter with words' spellings and thereby facilitate reading. In two different studies, they trained fourth-grade native English children on a set of novel words whose phonology made their spellings either highly predictable (e.g., /neʃ/ spelled as <nesh>) or highly unpredictable (e.g., /kɔɪb/ spelled as <koyb>). Children were first taught the pronunciations (phonological training) and the meanings (semantic training) of the words without seeing them in print. Next, they were tested reading those same words (hereafter trained words) embedded in sentences while their eye movements were measured. Children were also presented with sentences containing untrained novel words with both predictable and unpredictable spellings. As the authors expected, there was a significant facilitation effect for trained words with predictable as compared to those with unpredictable spellings. This facilitation was observed for all four eye-tracking measures, yielding shorter *total reading times*, shorter *gaze durations*, shorter *first fixation durations*, as well as fewer *regressions in* for words with predictable spellings. Importantly, there were no differences in reading words with predictable and unpredictable spellings in the group of untrained words. The authors interpreted this interaction between spelling predictability and training as evidence that children had generated orthographic expectations for all the words they previously acquired through phonological and semantic training. Since the expectations children had generated for predictable trained words matched their real spellings, the subsequent reading was facilitated. By contrast, there was a mismatch between children's orthographic expectations for unpredictable trained words and their actual spellings, such that no training facilitation emerged (see Beyersmann et al., 2021, for similar results observed in skilled adult readers).

The results presented by Wegener and colleagues provide strong evidence for the orthographic skeleton hypothesis, demonstrating that orthographic representations are generated solely from aural exposure to novel words. Importantly, given that facilitation was observed only for trained words with predictable spellings, the authors could argue that orthographic representations for novel words had been generated prior to readers' first visual encounter with the actual spellings and were not simply decoded during that first visual encounter. However, it remains unclear if the absence of processing facilitation (i.e., longer reading times) they observed for words with unpredictable spellings occurred because children generated inaccurate orthographic representations or rather because they did not generate any representations at all. Children may not have even initiated the process of generating orthographic expectations when uncertainty regarding potential spellings was high. A more direct test adjudicating between these two possibilities would be to train participants on a set of novel words controlled for the number of alternative spellings. The idea being that, if a word has only one possible spelling, all participants should generate the same orthographic representation. By contrast, if a word has multiple potential orthographic representations, it is difficult to predict, on a participant level, whether orthographic representations would be generated, and if so, which spelling option would be used to generate such representation. Indeed, Johnston et al. (2004) pointed out that the spellings they presented in their task might not have coincided with those their participants had imagined. But it seems clear that if participants generate certain orthographic expectations even under uncertainty (i.e., when there are multiple possible spelling options), they will still tend to generate a unique specific representation (selected from the possible options). This would be their preferred spelling and the one likely to be used to generate the orthographic skeleton.

All previous studies had been conducted in English, an opaque language with complex phoneme-to-grapheme conversion rules, in which both predictable and unpredictable words are likely to have more than one possible spelling. Therefore, manipulating the stimuli to include items with either a single or only two possible spellings would be challenging if not impossible due to both irregular spellings and complex phoneme-to-grapheme mappings (English contains 44 phonemes that map onto more than 200 graphemes). Spanish, by contrast, is a language with relatively simple phoneme-to-grapheme conversion rules (24 phonemes that map onto 32 graphemes). Most of the phonemes in Spanish map onto only one grapheme (e.g., sounds /p/ and /t/ can only be written as <p> and <t>, respectively), and almost all vowels are completely consistent. This means that there are many words (and pseudowords) whose spellings are entirely predictable from their phonology. These words (and pseudowords) have only one possible orthographic representation (e.g., the pseudoword /patu/ can only be written as <patu> in Spanish). At the same time, Spanish contains several inconsistent phonemes that have two orthographic representations (i.e., graphemes). For example, the sound /b/ can map either onto the letter <b> or the letter <v>. Similarly, the sound /χ/ when followed by vowels /i/ or /e/ can be written as either <g> or <j>. Consequently, if only one sound in a particular word is inconsistent, that word would have exactly two legal spellings (e.g., /χepo/ can be spelled as <gepo> or <jepo>). This property of the Spanish language provides a methodologically precise way of controlling the predictability of novel word spellings: creating words with either only one or only two possible spellings. In addition to comparing aurally trained and untrained words, this makes it possible to test whether orthographic skeletons are always generated (for both consistent and inconsistent words), regardless of any uncertainty related to phoneme-to-grapheme mappings, or are generated only when there is no uncertainty (i.e., only for consistent words, such as /patu/).

1.1. The current study

The aim of the present study was to test whether listeners generate orthographic skeletons for all words they acquire through aural instruction or whether they do so only when their expectations are likely to match the real spellings of those novel words (i.e., only for consistent words, which have one possible spelling). To that end, we created three groups of novel words that varied in terms of the number of possible spellings. Words from the *consistent* group comprised consistent phonemes only. As a result, these words had only one possible spelling, which was completely predictable from their phonology (e.g., /patu/ can only be written as <patu> in Spanish). Words from the second and third groups were inconsistent and had two possible spellings (e.g., /χepo/ can be spelled both as <gepo> and <jepo>).

To determine which of the two spellings would be preferred at the participant level, we obtained individual spelling preferences 2 weeks before the aural instruction took place. This was done to make sure we assess each participant's preferred and hence predictable spelling option. Consequently, words from the second group, the *inconsistent preferred* group, were always shown to participants in their preferred spellings, while words from the third group, the *inconsistent unpreferred* group, were invariably presented in each participant's unpreferred spellings. Since word spelling is known to depend strongly on the graphotactic rules of a language (Carrillo & Alegría, 2014), in some cases, the preferred spelling could be predicted solely based on bigram frequencies (i.e., one of the two possible grapheme representations is more likely to appear in a certain context; for instance, the grapheme <g> represents the sound /χ/ more frequently when followed by /i/). Some Spanish grapheme representations, however, have more balanced frequencies (e.g., the sound /b/ followed by vowels /a/ or /o/ can be written as either <b> or <v>). This makes it difficult to anticipate, at the group level, which spelling would be preferred between the two options. To account for individual differences in spelling preferences, we opted to assess them beforehand. We assumed that if participants generated unique orthographic skeletons for inconsistent phonologically trained words, they would be based on individual preferences (i.e., the spelling option they had provided beforehand). We determined participants preferred (and hence, likely to be predicted) and unpreferred (unlikely to be predicted) spellings through a pseudoword spelling task conducted 2 weeks before the main experiment.

Predictions regarding the outcomes of the study were the following. First, if participants generate orthographic expectations even when a word has more than one possible spelling, similar reading times should be observed for consistent and inconsistent preferred (i.e., likely to be predicted) trained words. Words from the inconsistent unpreferred group should elicit longer reading times because of the mismatch between the expected and the real spelling. Importantly, a comparison with the same three groups of untrained words should yield a significant interaction between training and word group. This interaction could be driven by a facilitation present when reading trained as compared to untrained consistent and inconsistent preferred words. Alternatively, the interaction could stem from significantly longer reading times present only for inconsistent unpreferred trained words (a surprisal effect). By contrast, if participants generate orthographic expectations only when these expectations are bound to match the real spellings, reading times should be faster for consistent (e.g., /patu/) than for inconsistent (e.g., /χepo/) trained items. Furthermore, no significant differences in reading times should be observed between the two inconsistent groups of words. In this case, the comparison with untrained words should lead to an interaction driven either by a facilitation present only for consistent-trained words or by longer reading times observed for all inconsistent-trained words (both preferred and unpreferred). Finally, no differences between the three groups of words should be observed in the set of untrained words since in this case, no orthographic expectations could have been generated prior to the first visual encounter. Untrained items should in addition serve as a control showing that any possible differences between three groups of words arose from phonological training only and were not a product of stimuli properties.

Table 1
Summary of objective (age of acquisition, picture naming, LexTale and interview) and subjective (self-rated) measures of participants' proficiency in Spanish

	Mean	*SD*	Range
Age of acquisition	0.00	0.00	0–0
Picture naming (BEST; 0–65)[a]	64.7	0.54	63–65
LexTALE (0%–100%)	93.0	6.24	71.7–100
Interview (1–5)	5.00	0.00	5–5
Self-rated proficiency (0–10)[b]			
Speaking	9.65	0.64	7–10
Understanding	9.64	0.61	8–10
Writing	9.45	0.77	7–10
Reading	9.51	0.75	7–10

Note. Some participants had some knowledge of a second or even a third language. However, none of them was highly proficient in any language other than Spanish.

[a]There are a total of 65 pictures to be named in the BEST (making 65 the maximum possible score).

[b]Self-rated proficiency data are missing for one participant.

2. Methods

2.1. Participants

A total of 54 participants completed the first session of the experiment. However, due to technical issues, three of them could not complete the second session. Moreover, due to low accuracy in the phonological training phase, additional three participants were excluded from any further analysis. Data reported here come from the 48 participants (44 females; *M*age = 25.6, *SD* = 3.74) who completed both experimental sessions within 14 to 16 days. All participants were native monolingual speakers of Spanish. Their language skills were assessed through a series of objective proficiency measures: An interview conducted by a native Spanish speaker rated from 1 (lowest level) to 5 (native or native-like level), a picture naming task (the BEST proficiency test; de Bruin, Carreiras, & Duñabeitia, 2017), a lexical decision task (i.e., LexTALE-Esp, which is the Spanish version of the LexTALE language proficiency test; Izura, Cuetos, & Brysbaert, 2014). Additionally, subjective measures of proficiency were obtained through participants' self-reports on different aspects of proficiency, such as writing, listening, understanding, and speaking (see Table 1). All participants were recruited from the internal BCBL database, and each received 15 euros for their participation in the study. The experiment was entirely web-based, but all participants had previous experience in participating in psychological experiments in the laboratory and were hence familiar with procedures and tasks used in experimental psychology. The experiment was approved by the BCBL Ethics Review Board (approval number 060420MK) and complied with the guidelines of the Helsinki Declaration. All participants gave their written consent at the beginning of each experimental session.

Table 2
The two sets of novel words used in the experiment

Set	Consistent	Inconsistent Preferred	Inconsistent Unpreferred
A	/dalu/	/ʎedu/	/ʎefo/
	/duti/	/ʎuɲe/	/ʎupo/
	/femi/	/χepo/	/χede/
	/fipu/	/χifo/	/χitu/
	/ludi/	/bamu/	/badi/
	/nepo/	/bupe/	/bumi/
	/panu/	/kime/	/kifo/
	/muni/	/ketu/	/keli/
B	/dopu/	/ʎepo/	/ʎeli/
	/fadi/	/ʎule/	/ʎufi/
	/leme/	/χeni/	/χetu/
	/mepu/	/χipe/	/χidu/
	/nute/	/bafu/	/bani/
	/pimu/	/buɲe/	/buti/
	/sufe/	/kipe/	/kiɲo/
	/tamu/	/kefi/	/kedi/

Note. Words from the inconsistent preferred group were later shown in each participant's preferred spelling, whereas words from the inconsistent unpreferred group were presented in the unpreferred spelling.

2.2. Stimuli

2.2.1. Novel words

Two sets of 24 four-phoneme-long (CVCV) bisyllabic novel words were created (Set A and Set B, later used for the trained and untrained items and counterbalanced across participants). Each set contained eight consistent and 16 inconsistent words. Consistent words were made up of consistent phonemes only, namely, those phonemes that map onto only one grapheme in Spanish. Consequently, consistent words had only one possible spelling (e.g., /dalu/ can only be written as <dalu>). Inconsistent words contained one of four target inconsistent phonemes that was always placed at the beginning of the word. All target inconsistent phonemes had two possible grapheme representations in Spanish: /b/ can be written as <b> or <v>; /k/ before vowels /i/ or /e/ can be written as <qu> or <k>; /ʎ/ maps onto <ll> or <y>; and /χ/ before vowels /i/ or /e/ can be written as <j> or <g>. Thus, all inconsistent words had two possible spellings in Spanish (e.g., /ʎedu/ can be written either as <lledu> or <yedu>) and both spellings adhered to Spanish orthographic rules. Inconsistent words were then split in half and assigned to either the preferred or unpreferred spelling group, which were based on individual productions collected in the pseudoword spelling task (eight words per group). Both groups contained two words starting with each of the four inconsistent target phonemes. As a result, both groups contained words that started with the following eight syllables: /ʎe/ and /ʎu/, /χe/ and /χi/, /ba/ and /bu/, /ki/ and /ke/ (see Table 2).

Consistent word sets were matched for number of orthographic neighbors (neither set had words with more than three neighbors, Set A: $M = 0.750$, $SD = 1.04$ and Set B: $M = 1.38$,

$SD = 1.22$, $t(14) = 1.12$, $p = .281$). We also ensured that both possible spellings of inconsistent items had no more than three orthographic neighbors. Note that none of the words ended with the vowel /a/, as this vowel generally marks feminine gender in Spanish nouns. This was done in order to avoid gender mismatch with the masculine demonstrative pronoun “este” used as the first word in all test sentences in the self-paced reading task in Session 2 (see Section 2.3.3). Participants were also explicitly told that all the novel words were masculine (see Section 2.3.2), discouraging them from imagining a word was feminine only to find it later presented with the masculine demonstrative pronoun “este.” Words were recorded by a male native Spanish speaker coming from the same region of Spain as the participants who took part in the study.

2.2.2. Novel objects

Forty-eight pictures from The Novel Object and Unusual Name (NOUN) Database (Horst & Hout, 2016) were selected and used as the novel objects participants were trained on (see Fig. 1). Pictures were divided into two sets of 24 pictures each (Set A and Set B) that were later, along with their associated words, counterbalanced across participants (half of the participants were trained on Set A and the other half on Set B). That is, one set was used for *trained items* and the other for *untrained items*. Pictures were randomly assigned names from the set they belonged to and were then kept constant for all participants.

2.3. Procedure

The entire study was done online using OSWeb online runtime, a JavaScript implementation of OpenSesame 3.3.2 software (Mathôt, Schreij, & Theeuwes, 2012). The experiment was hosted on and presented to participants through JATOS testing server (Lange, Kühn, & Filevich, 2015). All participants completed two experimental sessions with a 2-week pause in between. Both sessions started with the same audio message telling participants to use headphones throughout the experiment and adjust the sound to a comfortable level before initiating the experiment.

In Session 1, participants first completed a pre-test pseudoword spelling task in which they heard 96 pseudowords (half were target novel words and half filler pseudowords) and were instructed to spell them as if they were real words in Spanish. Next, they completed two linguistic distractor tasks (i.e., a lexical decision task followed by a real word spelling task). The aim of these two tasks was to mask the preferred spelling manipulation (i.e., make sure participants would not guess that the pseudoword spelling task was related to Session 2). Tasks were always presented in this order so as to avoid any orthographic effects from the distractor tasks on the pseudoword spelling task.

Two weeks after Session 1, participants received a link for Session 2, which comprised a training and a testing phase. Participants first received phonological training on the names of 24 novel objects. Next, they did a short non-linguistic distractor task (a two-colored Simon task; Simon, 1969), which served as a distractor task between the training phase and the self-paced reading task. After the Simon task, which took them around 3 minutes to complete,

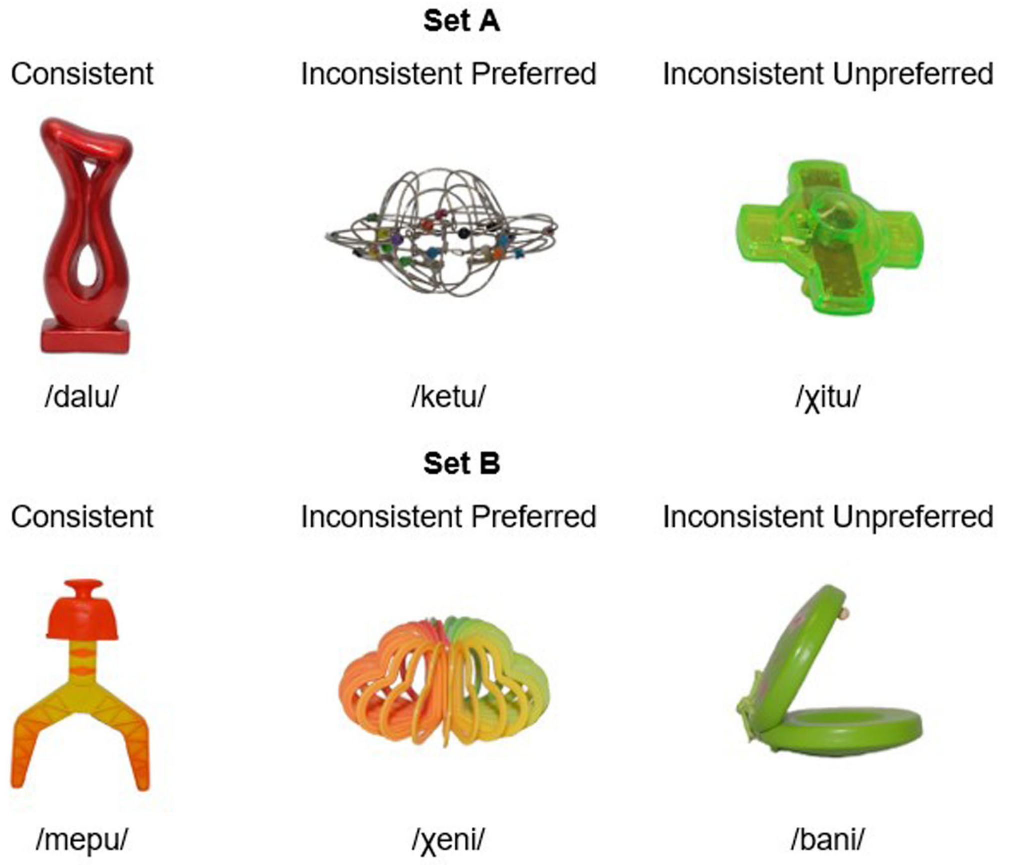

Fig. 1. An example object from each set (Set A and Set B) and word group (consistent, inconsistent preferred, and inconsistent unpreferred).

participants completed a self-paced sentence reading task in which they saw the written forms of words that they had or had not been trained on, embedded in short sentences.

2.3.1. Pseudoword spelling task

In the Session 1 pseudoword spelling task, participants heard a total of 96 pseudowords and were instructed to spell them as if they were real words in Spanish. Among these 96 pseudowords, 48 were target words (Set A and Set B; see Section 2.2.1) and 48 were filler pseudowords consisting of the same phonemes and first syllables as the target words (see Table A1 in Appendix A). The fillers were added to make sure participants would not remember, 2 weeks later during the phonological training session, that they had already heard all the novel words during the first session.

The aim of the pseudoword spelling task was to determine each participant's preferred spellings. These preferences were then used in Session 2 to determine the spellings of the items in the inconsistent preferred and inconsistent unpreferred group. For instance, if a

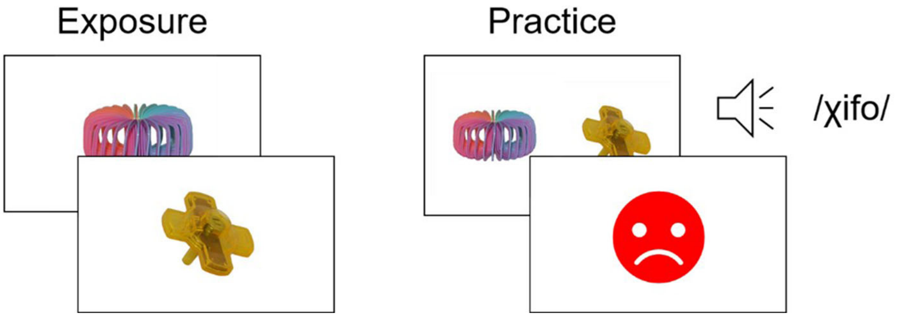

Fig. 2. The two parts of each block in the training phase. In the exposure part (left), participants saw each of the six objects from that block one by one, while listening to their names spoken three times in a row. In the practice part, they saw two objects on the screen and heard the name of one of them. Participants had to select the object that corresponded to the name they had heard by pressing either "M" (right) or "Z" (left) on their keyboard.

participant had preferred a spelling in Session 1, this spelling would be shown for a word from the inconsistent preferred group in Session 2 (e.g., if a participant spelled /kime/ with <k> in Session 1, that same participant would then see this word written with <k> in Session 2). Conversely, participants' not preferred spellings in Session 1 were used to present words from the inconsistent unpreferred group in Session 2 (e.g., if a participant wrote /kedi/ with <k> in Session 1, that same participant would then see the word written with <qu> in Session 2). Importantly, this was done for all 48 target words, that is, the 24 words participants were later trained on and the 24 untrained they had not been trained on.

Pseudowords were presented in a randomized order, and the task was self-paced. In each trial participants first heard a pseudoword over headphones. They were then prompted to spell the word that had just been played to them in the text box appearing below the question "Please spell the word you just heard". After typing in their response, they had to press "enter" to move on to the next trial and hear the next pseudoword. The task took participants around 10 minutes to complete.

2.3.2. Phonological training phase

During the phonological training phase (Session 2), participants were trained on the names of 24 novel objects belonging to one of the two sets (Set A or Set B). The sets were counterbalanced such that half of the participants were trained on Set A and the other half on Set B.

Before the task, participants were told that they would see pictures of some novel objects and that they should learn the names for a later test. Moreover, to internalize the masculine gender of the nouns during the training phase, participants were explicitly told that all the words they were about to hear were masculine (see Section 2.2.1).

In order to limit the training load, novel objects and their corresponding names were presented in four blocks of six novel objects/names. Each block contained two words taken from each of the three groups of words, yielding four blocks with identical structure. Moreover, each block was divided into two parts, an *exposure* and a *practice* part (see Fig. 2), and the order of the blocks was counterbalanced across participants. In the *exposure part*, participants saw pictures of six novel objects presented one by one at the center of the screen. While the picture was presented on screen, its name was played three times in a row at different speeds. The first and the third time, the name of the object was pronounced entirely, whereas the second time, it was pronounced by separating and emphasizing each of the two syllables (e.g., /muni/ → /mu/ – /ni/ → /muni/). Once participants had heard the name of the object three times in a row, they could press "enter" to continue and move on to the next trial. After seeing all six objects from one block, they proceeded to the *practice part*. In the practice part, two different objects appeared, one on the left and one on the right side of the screen while the name of one was played. Participants were prompted to select the object that corresponded to the name they heard by pressing "M" (for the picture on the right) or "Z" (for the picture on the left) on the keyboard. Each trial was immediately followed by feedback (a happy or sad face) that reinforced the training process. Each picture was paired with every other picture and appeared once on the left and once on the right side of the screen, for a total of 60 trials in each practice block. After completing each practice block, participants received feedback informing them of their overall accuracy rate (in %) in that practice block. This same procedure was repeated for all four practice blocks of words; participants were encouraged to take breaks between each block.

Once all the 24 objects had been presented and participants were trained to recognize their names, they completed the final check phase. This phase started with the exposure part, in which all 24 objects were presented once again, one by one at the center of the screen. At the same time, the name of each object was played through headphones only once. Participants moved from one picture to the next one by pressing "enter" on their keyboard. After being familiarized once again with the names of all objects, participants completed the final practice task. This time, they saw pictures of four objects on the screen, one on the left, right, top, and bottom areas of the screen (see Fig. 3). As in the previous practice phases, at the same time, the pictures were presented on the screen, participants heard the name of only one. They used the four corresponding arrows (left, right, up, down) on the keyboard to select the object on the screen corresponding to the name they had just heard. In order to make sure that each picture appeared the same number of times at each of the four positions on the screen and was paired an equal number of times with every other picture, position and pairing of the pictures was counterbalanced using a Latin square, giving a total of 144 trials. As in the previous practice phases, participants received feedback immediately after each trial, indicating whether their response was correct. They also received a final feedback message informing them about their overall performance at the end of the task. Only data from participants who obtained at least 70% of accuracy were considered in the main analysis (three participants were removed from the study based on this criterion; see Section 2.1). On average, participants took around 30 minutes to complete the entire phonological training.

2.3.3. Self-paced sentence reading task

In the self-paced sentence reading task (Session 2), participants saw the names of the 24 objects from the set they had been trained on (*trained words*), along with the names of the 24 objects taken from the other set (*untrained words*). All words, trained and untrained, were

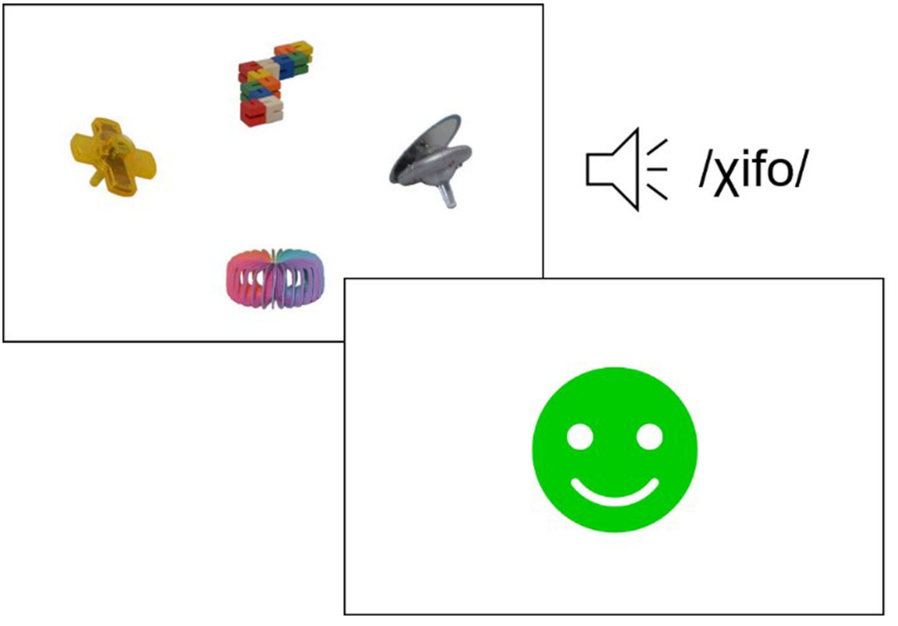

Fig. 3. Final practice phase at the end of the phonological training. In the final practice, participants saw four objects on the screen but heard the name of only one. Each trial was followed by a feedback message indicating whether their response was correct.

presented for the first time in writing in this task. The written names of the objects were embedded in eight different four-to-seven word sentences (see Table 3) that participants were asked to read word by word. Each target word appeared in each of the six possible positions in the sentence (two to seven), and thus the position of the target word in the sentence was counterbalanced across participants. This was done in order to show each target word an equal number of times in each sentence structure and at each position across all participants.

Table 3
Sentences from the self-paced reading task

Original Sentence	English Translation
Este **xxx** es pequeño	This **xxx** is small
Este gran **xxx** es bonito	This big **xxx** is pretty
Este es un **xxx** grande	This is a large **xxx**
Este es un pequeño **xxx** fantástico	This is a small fantastic **xxx**
Este objeto es un **xxx** pequeño	This object is a small **xxx**
Este objeto es un pequeño **xxx** bonito	This object is one small fantastic **xxx**
Este gran objeto es un **xxx** fantástico	This big object is one fantastic **xxx**
Este gran objeto es un fantástico **xxx**	This large object is a fantastic **xxx**

Note. The position of the target word is not equivalent in the Spanish sentences and their English translations due to syntactic differences across languages. Bold exes represent the place where target words appeared.

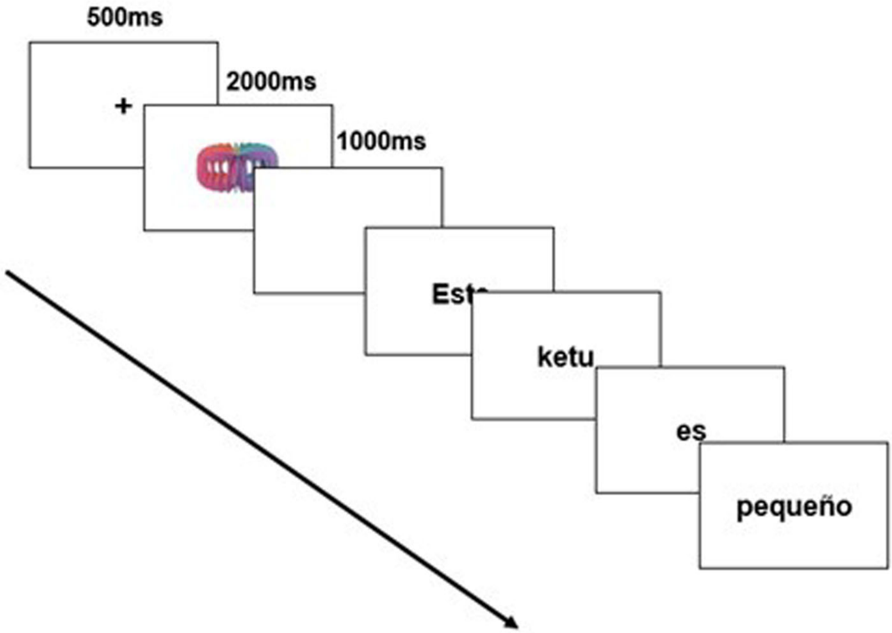

Fig. 4. The structure of the trial in the self-paced sentence reading task.

Varying the length of the sentences and the position of the target word in the sentence was done to avoid any anticipation of the target word reading (i.e., the moment of target word display within each sentence was unpredictable, ensuring that participants would not skip that word). To make sure all 24 words from each of the two sets were shown the same number of times in each position, a Latin square procedure was used, yielding eight different test orders. The eight sentences were thus repeated three times for the trained and three times for the untrained words (each participant read 48 sentences in total).

The picture of the object named in the sentence systematically preceded, thereby priming, the written presentation of the target word in the sentence. Sentences were presented in a randomized order, and each trial had the following structure: First, a fixation cross appeared at the center of the screen. After 500 ms, it was replaced by the picture of the object, which stayed on screen for 2000 ms. Next, a blank screen with a duration of 1000 ms was shown, after which the first word in the sentence appeared at the center of the screen. The demonstrative pronoun <este> was the first word in every sentence. Words were presented one by one, and the task was self-paced. Participants moved from one word to the next one by pressing “enter” on their keyboard. After reading the last word of the sentence, the participant initiated the beginning of the next trial by pressing “enter” (see Fig. 4). There were three practice sentences before the main task preceded by three known objects (e.g., a book, a glass, and a pencil). Participants were instructed to read each sentence as fast as possible without making pauses on any particular word. Reading latencies, that is, the time from the appearance of the target word at the center of the screen until participants pressed enter were recorded.

Table 4
Mean percentage of accuracy (*SD*s) per training block and in the final check phase

	Block1	Block2	Block3	Block4	Final Check
Set A	93.6 (4.61)	96.2 (4.77)	97.7 (2.61)	94.9 (5.09)	92.1 (8.62)
Set B	95.0 (4.91)	97.7 (2.72)	96.2 (4.20)	95.3 (4.49)	90.6 (7.45)

Participants with very short reading latencies, who were unlikely to have performed the task correctly, were excluded from final analyses. The same three participants who did not pass the phonological training session (i.e., who obtain less than 70% of accuracy in the final check phase) were also the only ones who had reading times ranging from 100 to 150 ms on almost all words. These participants were not included in the analysis. On average, participants needed around 10 minutes to complete the task.

3. Results

3.1. Pseudoword spelling task

The aim of the pseudoword spelling task was to obtain preferred spellings for each individual participant and hence make sure that preferred (likely to be predicted) and unpreferred spellings for inconsistent items were controlled for at the participant level rather than the group level. Preferred spellings per item are presented in Tables 6 and 7 in the Supplementary Material.

The target (first) phoneme was either misspelled or left out in 5.01% of all inconsistent items (words with two possible spellings). In cases where participants left out or misspelled a target sound (e.g., wrote "chuñe" pronounced as /tʃuɲe/, instead of "yuñe: or "lluñe" for the item /ʎuɲe/), their preferred spellings had to be inferred from the correct spellings of the filler words as well as other target items starting with the same syllable. More precisely, for each misspelled item, we looked at how all the other words starting with the same syllable had been spelled and selected the syllable used most often as the preferred one.

3.2. Phonological training phase

Overall accuracy in the final check in the phonological training was high: 91.4% ($SD = 8.01$, range 70%–100%), compared to an at chance level of 25%. Only five participants (among the 48 included in the final analyses) obtained less than 80% accuracy. Importantly, there were no significant differences in accuracy between the two sets of words (Set A: $M = 92.1$, $SD = 8.62$; Set B: $M = 90.6$, $SD = 7.45$; $t(46) = 0.627$, $p = .534$). Moreover, accuracy per training block (see Table 4) was also high, with an at chance level of 50%. None of the participants obtained less than 80% in any of the training blocks.

3.3. Self-paced reading task

Reading latencies for the target words (both trained and untrained) from the self-paced reading task were analyzed using linear mixed-effects models (Baayen, Davidson, & Bates, 2008) in the R statistical environment (Version 4.0.2; R Core Team, 2016). The analysis was performed using the *lme4* package (Version 1.1-23; Bates, Mächler, Bolker, & Walker, 2015), and all *p* values were obtained through *lmerTest* package (Version 3.1-2; Kuznetsova, Brockhoff, & Christensen, 2017).

Before analyzing reaction times (RT), extreme values (RTs below 150 ms and above 1200 ms) were identified by visual inspection and removed (4.9% of the data: no difference across conditions, $\chi^2 = (1, N = 113) = 3.17$, $p = .204$; Ratcliff, 1993; see also Baayen & Milin, 2010). To improve the positively skewed distribution, as well as minimize the effects of any possible outliers (Baayen, 2008), reaction times were log-transformed before the analysis. Log transformation was also in line with the Box–Cox test (Box & Cox, 1964).[2]

Given that planned comparisons were defined a priori (i.e., the inconsistent preferred group would significantly differ either from the consistent or from the inconsistent unpreferred group, but only in the set of trained words), a hypothesis-driven contrast coding approach was taken (see Schad, Vasishth, Hohenstein, & Kliegl, 2020). That is, a repeated contrast coding scheme was used to test the significance of the difference between the inconsistent preferred and consistent words against the difference between the two inconsistent groups of words. This coding scheme aims at testing differences between neighboring factor levels. In this particular case, the first contrast (hereafter, the group2-1 difference) tests the difference between inconsistent preferred and consistent words (consistent: −2/3, inconsistent preferred: 1/3, inconsistent unpreferred: 1/3). The second contrast (hereafter, the group3-2 difference) tests the difference between the two inconsistent groups of words (consistent: −1/3, inconsistent preferred: −1/3, inconsistent unpreferred: 2/3).

Fixed factor training was the first sum coded (trained words: 0.5, untrained words: −0.5), in order to look at the main effects of the two contrasts as well as their interactions with training. Next, to investigate differences of interest at each level of training (the group2-1 difference, and group3-2 difference for trained and untrained words only), two additional models were run with factor training was dummy-coded (i.e., in each of these two models, the level of interest, either trained or untrained words was set as a reference and coded as 0). Finally, to make sure there were no differences between the two sets of words, the set was sum-coded and added as a fixed covariate to the model (Set A was coded as 0.5 and Set B as −0.5).

To avoid overfitting the models which would lead to convergence issues and singular fits, and hence reduce statistical power (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), random effects structure was built through a stepwise model comparison procedure. Namely, the model including only random intercepts for participants/items was compared separately to several models with both by participants/items random intercepts as well as by participants/items random slopes for all experimental manipulations of interest. If adding a random effect term significantly improved model fit, it was included in the final model. Consequently, the reported models include the maximal random effects structure justified by the data.[3]

Table 5
Fixed and random effects structure of the model looking at the trained words

Fixed effects	β	*SE*	*t* value	*p*
(Intercept)	5.96	0.048	124	0.00***
Training	−0.038	0.017	−2.22	0.031*
Group2-1	0.006	0.024	0.241	0.810
Group3-2	0.063	0.024	2.56	0.010*
Set	0.146	0.096	1.53	0.134
Training: group2-1	0.014	0.034	0.401	0.688
Training: group3-2	−0.083	0.035	−2.40	0.016*
Random Effects	Variance	*SD*		
Participant: (intercept)	0.106	0.326		
Participant: Training (slope)	0.004	0.066		

Note. Factor training was dummy-coded (trained words coded as 0) to look at group2-1 and group3-1 differences only at the level of trained words. The two interactions, however, take both trained and untrained words into account.

*denotes statistical significance: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

The overall model (i.e., the model with training sum-coded) showed no main effects of either group2-1 difference ($\beta = 0.013$, $SE = 0.017$, $t = 0.743$, $p = .457$) or the group3-2 difference ($\beta = 0.021$, $SE = 0.017$, $t = 1.22$, $p = .223$). The main effect of training was significant ($\beta = 0.038$, $SE = 0.017$, $t = 2.24$, $p = .030$), showing that untrained words overall ($M = 419$, $SD = 144$) were read faster than trained ones ($M = 432$, $SD = 140$). Importantly, while the interaction between the group2-1 difference and training was not significant ($\beta = -0.014$, $SE = 0.035$, $t = -0.399$, $p = .689$), there was a significant interaction between the group3-2 difference and training ($\beta = 0.083$, $SE = 0.035$, $t = 2.4$, $p = .016$).

The model looking at the trained words only (trained coded as 0 and untrained as 1, see Table 5), showed that the group3-2 difference of 32 ms was statistically significant ($\beta = 0.063$, $SE = 0.024$, $t = 2.56$, $p = .010$), whereas the group2-1 difference of 10 ms was not significant ($\beta = -0.006$, $SE = 0.024$, $t = -0.241$, $p = .810$).

The same model with untrained words as a reference (untrained coded as 0 and trained as 1) showed that neither the group2-1 difference of 9 ms ($\beta = 0.020$, $SE = 0.024$, $t = 0.808$, $p = .419$) nor the group3-2 difference of −14 ms ($\beta = -0.020$, $SE = 0.025$, $t = -0.833$, $p = .405$) were significant.

To summarize, while the difference between consistent and inconsistent preferred trained words was not significant, the two groups of inconsistent trained words (preferred and unpreferred) differed significantly from each other. No differences were found for the untrained words (see Fig. 5).

3.3.1. Analysis based on bigram frequencies

To check whether bigram frequency plays a role in generating orthographic skeletons, the same models were run taking into account the bigram frequencies of the target syllables rather

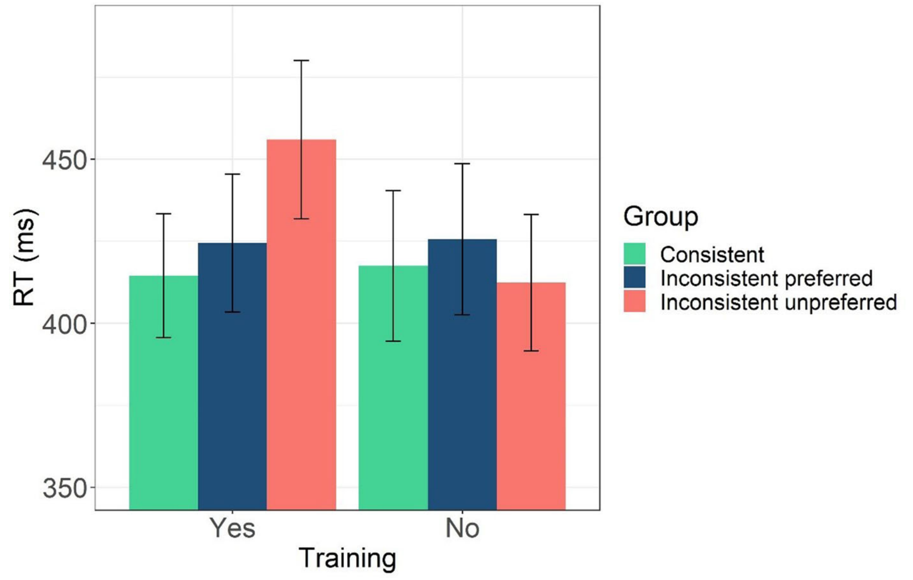

Fig. 5. Reaction times for all three groups of both trained (yes) and untrained (no) words. Error bars represent the standard error of the mean.

than participants' individual preferences. Namely, inconsistent words were divided into two groups, that is, they were classified as preferred or unpreferred, based on bigram frequencies calculated through B-Pal (Davis & Perea, 2005). This way, all words starting with bigrams *<ba>, <bu>, <je>, <gi>, <lle>, <llu>, <que>*, and *<qui>* were classified as preferred given that they are more frequently present in the initial position of real Spanish words. Their counterparts (i.e., less frequent initial bigrams) were hence considered as unpreferred.[4]

The model with trained words as the baseline showed no significant differences between consistent and preferred words ($\beta = 0.038$, $SE = 0.024$, $t = 1.55$, $p = .122$) nor a difference between the two inconsistent groups of words ($\beta = 0.017$, $SE = 0.025$, $t = 0.689$, $p = .491$). The same model with untrained words as the baseline also failed to show significant difference between consistent and preferred words ($\beta = 0.022$, $SE = 0.025$, $t = 0.897$, $p = .369$) or between the two inconsistent groups of words ($\beta = -0.009$, $SE = 0.025$, $t = -0.385$, $p = .700$). Importantly, there were no significant interactions between training and group2-1 difference ($\beta = -0.015$, $SE = 0.035$, $t = -0.444$, $p = .657$) or training and group3-2 difference ($\beta = -0.027$, $SE = 0.035$, $t = -0.765$, $p = .444$). Therefore, the pattern of results observed when considering participants' personal preferences was not replicated when preferred, and unpreferred spellings were inferred from statistical properties of the language (i.e., the bigram frequency).

4. Discussion

The goal of the present study was to further explore the orthographic skeleton hypothesis by testing whether people always generate orthographic expectations when acquiring novel spoken words, or only generate them when there is a unique possible spelling. Namely, the study aimed to investigate whether adult Spanish speakers generate preliminary orthographic representations for all words they acquire through aural exposure or only do so when it is certain that their expectations will match the real spellings of those novel words. To that end, 48 adult Spanish native speakers completed a two-session online experiment. In Session 1, individual spelling preferences were determined for each participant and for all novel words (i.e., both trained and untrained) from a pseudoword spelling task. In Session 2 (2 weeks later), participants were trained on the pronunciations of novel words pertaining to three groups (consistent, inconsistent preferred, and inconsistent unpreferred). Following this phonological training, participants were presented with both trained and untrained words in a self-paced sentence reading task. Overall, results suggest that Spanish adult readers generated orthographic expectations as a result of phonological training. Importantly, they did so for both consistent and inconsistent words (i.e., when there was certainty regarding spelling but also when there was uncertainty due to inconsistent phonemes). Furthermore, their orthographic skeletons for inconsistent items were in line with their individual spelling preferences, as inconsistent words shown in their preferred spellings did not differ in reading times from words with a single possible spelling (i.e., consistent words). Importantly, given that no differences were observed in the set of untrained words, and that slowing down for inconsistent unpreferred spellings occurred only in the group of previously acquired, that is, trained words, thus yielding a significant interaction between training and spelling, we take this as evidence that these effects indeed arise from orthographic expectations generated during phonological training.

The current study includes two important innovations: We controlled the number of spelling options for each novel word and used participants' personal spelling preferences in our critical manipulation. Given that predictable and unpredictable English words used by Wegener and colleagues could not be matched on the number of letters as well as bigram frequency, differences between the two groups of items might be at least partly linked to stimuli properties rather than orthographic expectations generated during the learning phase. This caveat was due to the complexity of phoneme-to-grapheme mappings present in the English language. By using a different language (Spanish) in our study, we could create words with either only one or two possible spellings. Furthermore, we were able to address an issue raised by Johnston et al. (2004) regarding individual variability in orthographic expectations by individually ascertaining each participant's spelling preferences, and hence the orthographic skeleton they were likely to generate. This proved to be a good strategy as preferred spellings varied considerably across participants (see Supplementary Material) and, in some cases, even deviated from the orthotactic rules of the language. For instance, based solely on the frequency of its appearance in Spanish, the grapheme <ll> should be preferred when writing words starting with the /ʎ/ sound. However, this was not the case as the majority of our participants preferred the grapheme <y>. Furthermore, as indicated by the absence of significant differences between the two inconsistent groups when the items were divided into preferred and unpreferred groups by initial syllable bigram frequency (i.e., the orthotactic rules of the language), personal preferences were indeed favored in generating preliminary orthographic representations. Finally, although it allowed us to adapt the stimuli material at the participant level, it can be argued that asking our participants to spell out all target words (both trained and untrained) could have influenced their performance on the task 2 weeks later. However, the precautions we took—including additional filler tasks and filler items to minimize the impact of the pre-test spelling task on the subsequent phonological training, as well as including a 2-week delay between the sessions—seem to have been enough to mask any potential influence of the pre-test task, as significant effects were present only in the group of trained words. Therefore, studies dealing with novel word spellings could, among other important psycholinguistic variables, consider adapting stimuli by taking participants' individual spelling preferences into account.

4.1. Comparison with Wegener et al. (2018)

Taken altogether, the present data support the orthographic skeleton hypothesis and show that its conclusions are valid even in a fairly transparent language. Additionally, by controlling for the number of alternative spellings, the study adds to the existing literature by showing that orthographic skeletons are generated even for words with more than one orthographically legal spelling. Finally, our results suggest that orthographic skeletons can be generated quickly, after relatively short phonological training including only an object's picture as semantic context. However, some important differences between the present study and the two studies conducted by Wegener and colleagues should be mentioned. First, the interaction between predictability (i.e., preferred spelling) and training observed in the present study stems from the longer reading times observed only for inconsistent trained words shown in their unpreferred spellings. However, the same interaction observed in their studies was driven by the facilitation observed only for trained words shown in predictable spellings. This reversed pattern of results, namely, the absence of an overall training advantage in the present study, could be explained by several differences between these studies: Wegener and colleagues used novel word learning paradigms that included both phonological and semantic training that was more extensive than the one employed in the present study. Furthermore, in their study, novel word learning took place over two experimental sessions. By contrast, participants in the present study went through a relatively short phonological training with only picture of the object as semantic context, and immediately after, in the same experimental session, were tested on reading these novel words. Moreover, different techniques were employed, and consequently different dependent variables were measured in these studies. While the conclusions of the present study are based on reading latencies measured through a behavioral response (i.e., button press after reading a word), Wegener and colleagues employed an online measure of the reading process (i.e., eye-tracking). The two studies also used different languages with highly distinct writing systems. English, which was used in all previous studies, is a highly inconsistent language with both phoneme-to-grapheme as well as grapheme-to-phoneme inconsistencies. Both reading as well as spelling out unfamiliar words in English involves high uncertainty (Ziegler, Stone, & Jacobs, 1997), and English readers are frequently confronted with unexpected spellings. This could lead to the facilitation effects observed when their expectations are confirmed. By contrast, Spanish readers, who are rarely confronted with inconsistencies and unexpected spellings, may be more sensitive to situations in which their expectations are not confirmed (as indicated by longer reading times only for unpreferred spellings). Finally, given the nature of our main task in which skilled adult readers were presented with short words embedded in relatively simple sentences with no semantic context (see Section 2.3.3), it is likely that the length of the target items (i.e., bisyllabic pseudowords) compromised the likelihood of detecting the training facilitation observed in the previous aural training studies (e.g., Álvarez-Cañizo, Suárez-Coalla, & Cuetos, 2019; Johnston et al., 2004; McKague et al., 2001). Therefore, apart from different techniques, different paradigms and designs also partly explain why there was no overall processing advantage for trained words in the current study.

Nevertheless, despite these differences, both studies reached the same important conclusion, namely, that orthographic representations of novel spoken words are generated prior to the first visual encounter with their real spellings. Importantly, these effects seem to be very robust as they have been observed in different populations (children as well as adult readers) and languages with different writing systems (opaque as well as transparent orthographies) and can be detected using different techniques (web-based as well as eye-tracking studies). Interestingly, our data show that web-based behavioral measures are sensitive enough to measure the orthographic skeleton effect, offering greater possibilities for testing this paradigm in more populations and in different languages. Indeed, the fact that evidence for the orthographic skeleton hypothesis was found even employing a single word presentation method, a noisier measure than the previously employed ones (i.e., eye-tracking and masked priming), speaks in favor of the hypothesis and its generalizability.

Therefore, these findings showing that visual forms of novel words are generated even in the absence of visual input during word pronunciation training (i.e., during the phonological training phase), have important implications for current models of reading development and visual word recognition. The new additional finding that orthographic skeletons are generated even when there is a risk of being wrong (i.e., when the spelling is uncertain) highlights the importance and prevalence of reading in our lives as well as our tendency to link orthography to spoken language.

On the whole, findings from the present study are in line with a broader line of psycholinguistic research showing the persistent effects of orthography on spoken word processing (Chéreau et al., 2007; Pattamadilok, Morais, Ventura, & Kolinsky, 2007; Rastle, McCormick, Bayliss, & Davis, 2011; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998). Once we learn to read, and more precisely, once we acquire phoneme-to-grapheme conversion rules, orthography inevitably affects word processing, even in the auditory modality. Previous research has shown that participants unconsciously access their orthographic knowledge even when performing tasks in which relying on orthography tends to be disadvantageous (Castles, Holmes, Neath, & Kinoshita, 2003; Tyler & Burnham, 2006). Similarly, written forms of already familiar words influence how fast we aurally recognize them (Chéreau et al., 2007; Seidenberg & Tanenhaus, 1979; Ventura, Morais, & Kolinsky, 2007; Ziegler & Ferrand, 1998). Our new findings supporting the orthographic skeleton account add to the previous psycholinguistic literature by showing that orthographic knowledge leads listeners to form preliminary orthographic representations of novel spoken words even in the absence of orthography.

4.2. Limitations and future directions

Before concluding, we should note some limitations of the current study. First of all, we based our findings on the assumption that any effects of phonological training on subsequent word reading arose as a result of orthographic representations that were generated during phonological training and retained in lexical memory. However, it is also possible that these orthographic representations were not retained in lexical memory but instead activated during the first moment of visual encounter with the word's spelling. That is, when confronted with the written form of the word (e.g., <vamu>), a phonological representation of the word might have been automatically generated. This phonological representation would match the one created during the training phase, which would in turn activate its preferred orthographic form. This later representation would match or mismatch the orthographic word currently being presented and hence facilitate or not facilitate word recognition. It is important to note, however, that the predictions that result from this alternative hypothesis are exactly the same as those implied by the orthographic skeleton hypothesis. Future research could set out to adjudicate between the two possibilities. Second, our results showed that orthographic expectations are generated even when there is uncertainty regarding the possible spellings (i.e., even for inconsistent items). However, the inconsistent words in this study had only two possible spellings that differed only on a single phoneme-to-grapheme mapping. As the number of inconsistent phonemes—and thus the number of possible word spellings—increases, it is possible that people might stop generating orthographic skeletons due to the higher probability of a mismatch between their orthographic expectations and the actual spellings. Alternatively, they might always select a single spelling based on their personal preferences. Therefore, the ubiquity of skeleton creation when learning novel words should be explored in other languages with more opaque writing systems than Spanish. Another consideration is that skilled adult readers may be more impacted by orthography than children and may therefore be more likely to generate orthographic skeletons for all novel words. By contrast, early and developing readers might generate orthographic skeletons only when a single highly probable spelling is available (only for consistent words) due to less experience with orthotactic probabilities. For this reason, exploring the skeleton hypothesis in a sample of Spanish children would help us determine the extent to which the orthographic skeleton hypothesis can be generalized to different populations (i.e., early readers). Finally, the present study cannot answer the question as to whether these orthographic representations are generated unconsciously as an automatic response to any phonological input (Castles et al., 2003; Johnston & Castles, 2003; Tyler & Burnham, 2006) or instead serve as a mnemonic tool that helps participants learn novel words under experimental conditions (Johnston et al., 2004; Rosenthal & Ehri, 2008). Indeed, all previous studies employed word learning paradigms in which participants were explicitly instructed to learn novel words. While it is possible that orthographic skeletons were generated automatically and unconsciously upon hearing the novel words, it remains possible that participants consciously generated orthographic skeletons in order to support word memorization. Further studies comparing skeleton creation in word learning versus passive word listening contexts would help to address this question.

5. Conclusion

Previous research has shown that preliminary orthographic representations can be generated solely from aural exposure to novel words. Importantly, this happens prior to the first visual encounter with the actual spellings of those novel words. The present study provides further evidence for this account and adds to the existing literature by showing that participants generate orthographic expectations for novel words with two possible spellings according to their individual preferences, that is, despite uncertainty regarding their real spellings. The current research demonstrates that the orthographic skeleton account is valid even for a transparent orthography with few phoneme-to-grapheme inconsistencies and that spelling expectations can be generated quickly, even after relatively short phonological training.

Acknowledgments

The authors would like to thank Magda Altmann for her help with editing the article; Larraitz López, Amets Esnal, and Itziar Basterra for their help with data collection as well as the anonymous reviewers for their comments, which improved the quality of the article. This research was supported by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No: 819093 to CDM), the Spanish Ministry of Economy and Competitiveness (PID2020-113926GB-I00; PSI2017-82941-P to CDM) and the Basque Government (PIBA18_29 to CDM). The first author was supported by a predoctoral fellowship (associated with the project PSI2017 82941-P; Grant No: PRE-2018-083946) from the Spanish Ministry of Science, Innovation and Universities and the Fondo Social Europeo.

Conflict of interest

The authors declare that there is no conflict of interest.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data are available at https://osf.io/h69dg/?view_only=3b8804c6ccdf46c5b6ebe83c15e9965b and materials are available at https://osf.io/h69dg/?view_only=3b8804c6ccdf46c5b6ebe83c15e9965b.

Notes

1 We employ the term *orthographic expectations* to denote the orthographic representation of a word that has been constructed in the absence of its written form, that is, solely based on aural exposure to the word.

2 Note that the lower rate of outlier removal (e.g., < 1.5% of the data or even no removal at all) led to the same pattern of significance in the main analysis.

3 The exact structure of the model was the following: logRT∼ training*group2-1 + training*group3-2 + set + (1+ training||participant). Note that the final model did not include random intercepts for items, as including them indicated singular fits, meaning that the variance associated with these items was estimated as (close to) zero. Nevertheless, removing random intercepts for items did not change the pattern of significance, and in fact, including the full random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013) yielded the same results.

4 Note that distribution and number of items belonging to each of the two groups were not balanced across participants, as some items were classified as preferred for some participants, while for the others, those same items were unpreferred. For example, a word /bafu/, belonging to the preferred group, was presented to some participants as <bafu> and to some as <vafu> in the reading task (based on their spelling preferences). However, based on bigram frequency, this item was classified as preferred for all participants who saw it with the letter <b> in the task. By contrast, it was classified as unpreferred for those who saw it with <v>.

References

Álvarez-Cañizo, M., Suárez-Coalla, P., & Cuetos, F. (2019). Orthographic learning in Spanish children: Influence of previous semantic and phonological knowledge. *Journal of Research in Reading*, *42*(1), 137–149. https://doi.org/10.1111/1467-9817.12254

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beyersmann, E., Wegener, S., Nation, K., Prokupzcuk, A., Wang, H.-C., & Castles, A. (2021). Learning morphologically complex spoken words: Orthographic expectations of embedded stems are formed prior to print exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 87–98. https://doi.org/10.1037/xlm0000808

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Carrillo, M. S., & Alegría, J. (2014). The development of children's sensitivity to bigram frequencies when spelling in Spanish, a transparent writing system. *Reading and Writing*, *27*(3), 571–590. https://doi.org/10.1007/s11145-013-9459-y

Castles, A., Holmes, V. M., Neath, J., & Kinoshita, S. (2003). How does orthographic knowledge influence performance on phonological awareness tasks? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *56A*(3), 445–467. https://doi.org/10.1080/02724980244000486

Chéreau, C., Gaskell, M. G., & Dumay, N. (2007). Reading spoken words: Orthographic effects in auditory priming. *Cognition*, *102*(3), 341–360. https://doi.org/10.1016/j.cognition.2006.01.001

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*(4), 665–671. https://doi.org/10.3758/bf03192738

de Bruin, A., Carreiras, M., & Duñabeitia, J. A. (2017). The BEST dataset of language proficiency. *Frontiers in Psychology*, *8*, 522. https://doi.org/10.3389/fpsyg.2017.00522

Duff, F. J., & Hulme, C. (2012). The Role of Children's phonological and semantic knowledge in learning to read words. *Scientific Studies of Reading*, *16*(6), 504–525. https://doi.org/10.1080/10888438.2011.598199

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. https://doi.org/10.1037/0278-7393.10.4.680

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). *Lexical Test for Advanced Learners of Spanish (Lextale-ESP) [database record]*. Washington, DC: APA PsycTests. https://doi.org/10.1037/t47086-000

Johnston, M. B., & Castles, A. (2003). Dissociating automatic orthographic and phonological codes in lexical access and lexical acquisition. In S. Kinoshita, & S. J. Lupker (Eds.), *Macquarie monographs in cognitive science. Masked priming: The state of the art* (pp. 193–222). New York, NY: Psychology Press.

Johnston, M., McKague, M., & Pratt, C. (2004). Evidence for an automatic orthographic code in the processing of visually novel word forms. *Language and Cognitive Processes*, *19*(2), 273–317. https://doi.org/10.1080/01690960344000189

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

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. https://doi.org/10.1111/j.1467-9817.2007.00361.x

McKague, M., Pratt, C., & Johnston, M. B. (2001). The effect of oral vocabulary on reading visually novel words: A comparison of the dual-route-cascaded and triangle frameworks. *Cognition*, *80*(3), 231–262. https://doi.org/10.1016/S0010-0277(00)00150-5

Nation, K., & Cocksey, J. (2009). The relationship between knowing a word and reading it aloud in children's word reading development. *Journal of Experimental Child Psychology*, *103*(3), 296–308. https://doi.org/10.1016/j.jecp.2009.03.004

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*(4), 342–356. https://doi.org/10.1111/j.1467-9817.2004.00238.x

Pattamadilok, C., Morais, J., Ventura, P., & Kolinsky, R. (2007). The locus of the orthographic consistency effect in auditory word recognition: Further evidence from French. *Language and Cognitive Processes*, *22*(5), 700–726. https://doi.org/10.1080/01690960601049628

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain Research*, *1188*, 132–138. https://doi.org/10.1016/j.brainres.2007.10.084

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1588–1594. https://doi.org/10.1037/a0024833

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology*, *100*(1), 175–191. https://doi.org/10.1037/0022-0663.100.1.175

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 546–554. https://doi.org/10.1037/0278-7393.5.6.546

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. https://doi.org/10.1037/h0027448

Stuart, M., & Coltheart, M. (1988). Does reading develop in a sequence of stages? *Cognition*, *30*(2), 139–181. https://doi.org/10.1016/0010-0277(88)90038-8

Tyler, M. D., & Burnham, D. K. (2006). Orthographic influences on phoneme deletion response times. *Quarterly Journal of Experimental Psychology*, *59*(11), 2010–2031. https://doi.org/10.1080/17470210500521828

Ventura, P., Morais, J., & Kolinsky, R. (2007). The development of the orthographic consistency effect in speech recognition: From sublexical to lexical involvement. *Cognition*, *105*(3), 547–576. https://doi.org/10.1016/j.cognition.2006.12.005

Wegener, S., Wang, H. C., de Lissa, P., Robidoux, S., Nation, K., & Castles, A. (2018). Children reading spoken words: Interactions between vocabulary and orthographic expectancy. *Developmental Science*, *21*(3), 1–9. https://doi.org/10.1111/desc.12577

Wegener, S., Wang, H-C., Nation, K., & Castles, A. (2020). Tracking the evolution of orthographic expectancies over building visual experience. *Journal of Experimental Child Psychology*, *199*, 104912. https://doi.org/10.1016/j.jecp.2020.104912

Ziegler, J., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, *5*(4), 683–689. https://doi.org/10.3758/bf03208845

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. https://doi.org/10.3758/bf03210615

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table 6 Novel words from Set A
Table 7 Novel words from Set B

Appendix A

Table A1
Filler pseudowords from the pseudoword spelling task

Consistent	Inconsistent	Inconsistent
/nufa/	/besu/	/kika/
/lifa/	/biɲo/	/kodu/
/lusi/	/baru/	/kado/
/tado/	/beʧo/	/kebo/
/fasa/	/bugo/	/kesi/
/dofa/	/beli/	/kibe/
/nadu/	/boʧi/	/kigo/
/mita/	/bafa/	/keλo/
/nafo/	/λado/	/χibu/
/meli/	/λebi/	/χeko/
/teda/	/λoto/	/χeɲa/
/tefi/	/λubo/	/χeλa/
/mafe/	/λomo/	/χifa/
/lono/	/λaku/	/χega/
/puda/	/λepa/	/χigo/
/pefo/	/λuχa/	/χeru/