Morphology (2021) 31:275–296
https://doi.org/10.1007/s11525-020-09363-5

# Sensitivity to meaningful regularities acquired through experience

**Anastasia Ulicheva**[1] · **Marco Marelli**[2] · **Kathleen Rastle**[1]

Received: 25 September 2019 / Accepted: 23 June 2020 / Published online: 2 July 2020

**Abstract** English spelling provides multiple cues to word meaning, and these cues are readily exploited by skilled readers. In two crowdsourcing studies, we tested skilled readers' sensitivity to a large number of morphological as well as nonmorphological orthographic cues by asking them to classify nonwords as adjectives or nouns. We observed a substantial variability across individuals and orthographic cues. In this paper, we discuss some sources of this variation. Specifically, we found consistent correlations between readers' sensitivity to cues and their performance on language tasks (reading, spelling, and author recognition tests) suggesting that reading experience is critical for assimilating spelling-to-meaning regularity from written language. Further, we identified characteristics that may be important for the learning and exploitation of orthographic cues that are related to the nature of their function and use in context.

**Keywords** Spelling · Meaning · Morphology · Lexical category · Distributional semantics · Individual differences

## 1 Introduction

In alphabetic writing systems, letters primarily represent sounds. Groups of letters, however, occur repeatedly in words with similar meanings ("trust", "distrust", "trustworthy"), and thus form islands of regularity between form and meaning (Rastle et al. 2000). Most commonly, these islands of regularity are morphemes with transparent, compositional meanings. These morphemes can be stems (e.g., TRUST) or affixes (e.g., –LY) that alter the meanings of words in predictable ways (e.g.,

✉ A. Ulicheva
Ana.Ulicheva@rhul.ac.uk

[1] Department of Psychology, Royal Holloway, University of London, Egham, UK

[2] Department of Psychology, University of Milano-Bicocca, Milan, Italy

"wildly", "boldly", "kindly"; Rastle et al. 2000). Nonetheless, other letter groups that traditionally are not considered morphemes can also reoccur in different words, while carrying some degree of meaningful information: glisten, glass, glitter (has something to do with light); snore, sneeze, snort (has something to do with nose); cranberry, grocery (Harm and Seidenberg 2004; Rastle et al. 2000; Aronoff 1976; Seidenberg and Gonnerman 2000).

A long-standing debate in the psycholinguistic word recognition and reading research concerns the question of whether morphemes are explicitly represented in the reading system. For example, Taft and Forster (1975) demonstrated that nonwords with an apparent morphological structure (such as "dejuvenate" that consists of an existing prefix DE– and stem JUVEN as in "juvenile", "rejuvenate") take longer to reject in a lexical decision task compared to nonwords without such structure (e.g., "depertoire" whose "stem" does not exist; see also Rastle et al. 2000; Forster et al. 1987; Rastle et al. 2004). This line of evidence suggested that morphemes may enjoy a special processing advantage compared to matched nonmorphological patterns. However, an influential account of derivational morphology (Seidenberg and Gonnerman 2000; Harm and Seidenberg 2004) eschews a separate level of representation dedicated to morphemes. Morphemes and other regular or quasiregular patterns emerge as distributions across hidden units representing statistical regularities that hold across orthographic, phonological, and semantic information (Seidenberg and Gonnerman 2000). In such models, regularities that consistently hold across multiple levels (orthography, phonology, semantics) are particularly salient. This feature of the model can potentially account for the effects that have been considered to be purely morphological in nature (Taft and Forster 1975; Rastle et al. 2004).

Leaving those theoretical debates aside, all reading scholars would agree that the ultimate goal of the reading process is recovering word meaning from orthography. One aspect of word meaning that appears particularly important for successful comprehension is lexical category (e.g., 'we saw her duck' where "duck" may mean a noun or a verb, depending on a broader context). Cues to lexical category are remarkably salient in spelling. For example, there are many morphemes in English whose spellings are consistent across multiple occurrences in different words, whereas their pronunciations are not (Berg and Aronoff 2017; Rastle 2019; Ulicheva et al., 2020). For instance, English past tense verbs may end in /əd/, /d/ or /t/ phonologically, but these final phonemes are always spelled –ED (Carney 1994). Interestingly, such spelling cues are not always morphemic. For example, context and function words with the same pronunciation may take different spellings ("inn", "bye" vs "in", "by", Smith et al. 1982; Albrow 1972). Smith et al. discuss similar differences in final letter doubling to distinguish proper nouns (e.g., "Kidd", "Carr") from common nouns (e.g., "kid", "car").

### 1.1 Readers' sensitivity to morphological and nonmorphological cues

A large body of research suggests that both morphological and nonmorphological sources of information are actively used by readers for comprehension. Ulicheva et al. (2020) studied morphological cues to meaning. The authors conducted a computational analysis of English derivation showing that suffix spellings carry unique information about meanings (specifically, about lexical categories) that is not always available in the phonological forms of suffixes (see also Berg and Aronoff 2017). Ulicheva et al. (2020) designed a measure for the amount of this information present within spellings and labelled it “diagnosticity”. Diagnosticity refers to the number of words with a given suffix spelling that belong to a specific lexical category, divided by the total number of words with this suffix spelling. The authors showed that diagnosticity values of English suffixes were high, with the mean of 0.78 (diagnosticity values range from 0 to 1) indicating that English derivational suffixes are reliable markers of meaning. Further, in two behavioural experiments, Ulicheva et al. (2020) have shown that skilled readers possess the knowledge of this meaningful information and rapidly exploit it when they read. For example, in their Experiment 1, forty-six participants made noun/adjective category judgements to nonsense words such as “jixlet”. Ten noun and ten adjective suffixes that varied in diagnosticity were used to form nonwords. Overall, participants were more likely to classify nonwords as adjectives when they ended in adjective-diagnostic suffixes than when they ended in noun-diagnostic suffixes. Further, reading and spelling behaviour mirrored the strength with which suffix spellings cued category in a large corpus. In other words, as suffixes became more diagnostic for a given category, participants’ responses increasingly favoured that category. Based on these data, the authors suggested that skilled readers’ long-term knowledge represents the statistical structure of the writing system and that this knowledge is likely acquired through implicit statistical learning processes (see also St. Clair et al. 2010).

On the other hand, a large body of literature suggests that nonmorphological orthographic patterns (e.g. –OON that occurs predominantly in nouns, e.g., “noon”, “balloon”) also carry meaningful category information, and that skilled readers are able to exploit these alongside morphological cues. Kemp et al. (2009) demonstrated across three tasks that skilled readers are generally sensitive to nonmorphological letter sequences that are diagnostic for nouns (e.g., –OON) or verbs (e.g., –ERGE as in “diverge”, “emerge”; see also Arciuli and Cupples 2003, 2004, 2006; Farmer et al. 2006; Kelly 1992; Arciuli and Monaghan 2009; Cassani et al. 2020). One important aspect of this study is that they found that reading ability was correlated with cue sensitivity in sentence construction ($r$ =.32, $p$ =.008) and sentence judgement tasks ($r$ =.26, $p$ =.038). Kemp et al. reasoned that these correlations reflect a gradual build-up of meaningful (nonmorphological) information through repeated exposure to letter strings through reading (see also Rastle 2019; Farmer et al. 2015; Arciuli et al. 2012). Studies in adjacent domains support the view that the degree to which probabilistic patterns, such as print-to-sound correspondences, can be learnt from exposure depends on the richness of reading experience (Steacy et al. 2019; see also Treiman et al. 2006).

To our knowledge, there has been only one attempt to compare the processing of morphological and nonmorphological cues to meaning directly. Using MEG, Dikker et al. (2010) found differences in early visual cortex activity as early as 120 ms following exposure to category-typical and category-atypical words. Stimulus words contained endings that were either suffixes (as in “farmer”, “artist”) or nonmorphological endings (as in “movie”, “soda”). No differences were reported in people’s sensitivity to the two types of cues. However, as a closer inspection of their materials suggests, the nonmorphological status of some of the cues used by Dikker et al. is debatable. These potentially problematic words included those ending in –LE, –AR, –ESS, –IC, or up to 55% of all words in this condition, so the results of this study should be interpreted with caution.

## 1.2 Distributional characteristics of morphological and nonmorphological cues to meaning

Based on the studies reviewed above, there seems to be little reason or evidence to believe that morphological and monmorphological cues to meaning are exploited differently by readers, provided that these have similar distributions in a language. Nonetheless, prototypical affixes and nonmorphological orthographic patterns tend to have rather different distributional characteristics (Ulicheva et al. 2020). Figure 1 illustrates such differences in terms of diagnosticity and frequency. English suffixes and orthographic endings were extracted from CELEX. Orthographic patterns were defined as all one to five letter sequences that appeared word finally and were not endings of existing suffixes. It can be seen from Fig. 1 that many morphemes and endings are equally frequent and diagnostic, but a typical suffix is more frequent and more diagnostic than a typical nonmophological ending. These differences between the two ends of the spectrum may have implications for learning (Frost 2012). Due to their productivity, suffixes and the information that they carry may be easier to learn than most non-morphological endings (Tamminen et al. 2015). Further, suffixes are higher in diagnosticity than most non-morphological endings, which may boost their learnability (Tamminen et al. 2015). The third relevant difference is that most orthographic endings cue noun meanings (88%), while suffixes are primarily adjective (30%) and noun (65%) forming (Ulicheva et al. 2020).

In fact, there may be distributional differences between morphological and nonmorphological patterns that concern their mapping to meaning or their use in context that are less widely discussed. For instance, morphological information provides semantic detail that goes beyond mere category information (–ER means an agent, –ESS often corresponds to a female agent etc., see Seidenberg and Gonnerman 2000; Marelli and Baroni 2015). Generally speaking, such fine-grained consistent information that is encoded in morphological units might be more readily available for learning than that encoded in orthographic endings. Secondly, while affixes modify meanings of stems, not all do so in predictable, transparent ways (Marelli and Baroni 2015). These differences in the amount and type of content that units carry might also imply differences in usage: one might speculate that morphemes are used predictably in specific, semantically related contexts, while nonmorphological endings could be used more broadly, appearing in semantically unrelated words across a wider variety of contexts. Identifying the factors that influence the learning and exploitation of meaningful information is valuable for the refinement of existing models of reading.

In order to understand if there are any processing differences between morphological and nonmorphological orthographic patterns, we designed two online crowdsourcing experiments where skilled readers were asked to classify nonwords ending in morphological and nonmorphological cues that were matched on two important distributional characteristics (i.e., frequency and diagnosticity). The two experiments

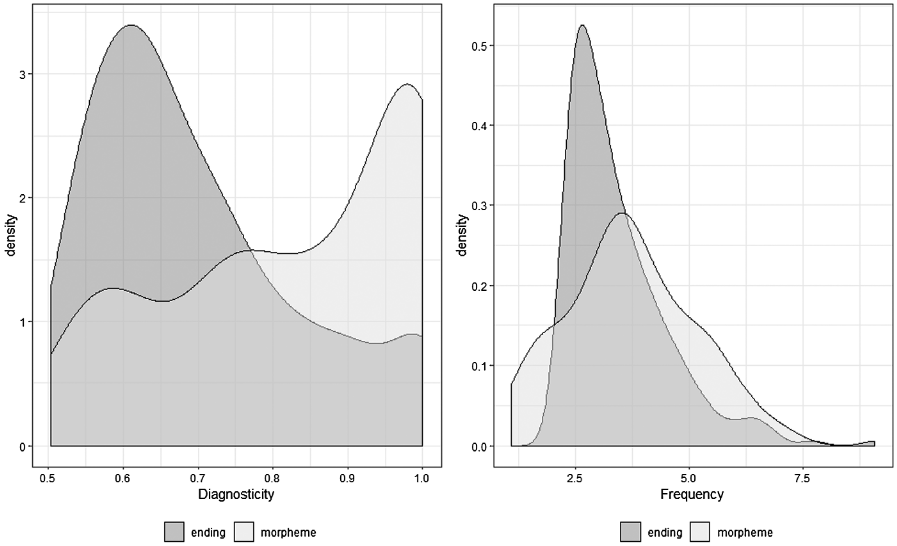

**Fig. 1** Distributional properties of orthographic endings and suffixes. Left panel is a density plot of diagnosticity. Right panel is a density plot of logarithm-transformed frequency values

**Table 1** Suffixes and endings for the experiments. Noun suffixes printed in italics were removed from the noun suffix/noun ending comparison, because matching nonmorphological endings were not available

| Adjective suffixes | Noun suffixes | Noun endings | Adjective suffixes | Noun suffixes | Noun endings |
|---|---|---|---|---|---|
| ile | ite | uff | ular | ster | ird |
| ant | ine | dge | etic | ition | ra |
| ary | i | ome | ish | ee | ork |
| y | ard | ush | ual | ery | olk |
| ern | eer | z | ic | age | iece |
| ory | ade | ob | atic | *er* | – |
| al | ey | que | ible | let | logue |
| ful | ure | od | like | ette | na |
| id | our | del | ical | ance | illa |
| ent | ice | ord | less | *ment* | – |
| ive | et | tre | able | *ness* | – |
| ial | ist | oom | ous | *ism* | – |
| ific | ety | um | | | |

differed in terms of the type of judgement participants had to make. Experiment 1 used an explicit task that required classifying isolated nonwords into adjective vs noun categories. Experiment 2 used an implicit category judgement task where participants had to decide how well nonwords fit into sentences. Our first question concerned item-based variability: are there actual processing differences between different types of cues, and is there any evidence on differences in the learning of this information? We hypothesised that people's sensitivity to suffix cues might be stronger than that to nonmorphological cues for the reasons outlined in the previous paragraph. For the purposes of the categorisation task, an equal number of adjective endings had to be included in the experiment. This provided us with an opportunity to compare people's behaviour towards noun suffixes vs adjective suffixes, although no differences between categories were expected *a priori*. An additional question here concerned the influence of diagnosticity on participants' responses. Graded effects of diagnosticity are interpreted as evidence for an involvement of statistical learning mechanisms in learning (Ulicheva et al. 2020). Therefore differential effects of diagnosticity on ending types serve as a window into understanding item-based variability, i.e., how different endings might be acquired. Finally, following Kemp et al. (2009), we were also interested in explaining any differences that might arise across individuals in sensitivity to both types of cues, and relating these differences to participants' language skills. In particular, we hypothesised that better sensitivity to cues would be associated with better linguistic ability, or with more reading experience. To this end, we expected to find better sensitivity to cue diagnosticity in participants with better spelling, vocabulary, and tests of reading experience.

## 2 Experiment 1

### 2.1 Method

#### *2.1.1 Materials*

Three types of letter endings were used in the main experiment: noun suffixes, adjective suffixes, and nonmorphological noun endings. We identified only four nonmorphological adjective endings (–IKE, –LETE, –UL, –UNG), and therefore were unable to take advantage of this manipulation. Two comparisons were planned: (1) that between noun and adjective suffixes, and (2) that between noun suffixes and noun endings. Table 1 lists all endings; Table 2 lists descriptive statistics for psycholinguistic variables. Suffixes and nonmorphological endings, i.e., endings of non-suffixed words, were extracted from CELEX (Baayen et al. 1993). Twenty-five noun suffixes were matched to 25 noun endings on type frequency, diagnosticity, and length in letters (see Table 2). Note that token frequency was not controlled in this experiment, and noun endings were lower in token frequency than noun suffixes ($t = 4.64$, $p < 0.0001$). The type diagnosticity measure captured the amount of meaningful information in a given spelling. For a particular spelling, diagnosticity is calculated by dividing the number of words ending in this spelling and falling into this category by the total number of words that contain the spelling (see Ulicheva et al. 2020, for details). Type frequency is the number of words in CELEX that ended in given letter patterns. For instance, the frequency value for –ER included pseudoaffixed words such as CORNER, as well as morphologically simple words such as ORDER.

Only 21 nonmorphological endings that could be matched to noun suffixes on frequency and diagnosticity were identified, because nonmorphological endings are typically characterised by substantially lower values on both metrics (see Fig. 1).

**Table 2** Descriptive statistics for matching variables (no suffix exclusions)

| | | Min | 1st quantile | Mean | Median | SD | 3rd quantile | Max |
|---|---|---|---|---|---|---|---|---|
| Diagnosticity | Adjective suffixes | 0.548 | 0.759 | 0.825 | 0.820 | 0.116 | 0.922 | 0.991 |
| | Noun suffixes | 0.550 | 0.734 | 0.822 | 0.857 | 0.132 | 0.916 | 1 |
| | Noun endings | 0.553 | 0.738 | 0.809 | 0.838 | 0.122 | 0.909 | 1 |
| Type frequency (logarithm-transformed) | Adjective suffixes | 2.300 | 3.34 | 4.30 | 4.53 | 1.21 | 5.32 | 6.08 |
| | Noun suffixes | 2.480 | 3.27 | 4.13 | 3.69 | 1.23 | 4.75 | 7.12 |
| | Noun endings | 2.400 | 3.09 | 3.57 | 3.66 | 0.71 | 4.14 | 4.92 |
| Token frequency (logarithm-transformed) | Adjective suffixes | 7.87 | 9.70 | 10.0 | 10.1 | 1.01 | 10.5 | 12.2 |
| | Noun suffixes | 7.58 | 9.39 | 10.2 | 10.6 | 1.02 | 11.7 | 11.7 |
| | Noun endings | 6.38 | 8.08 | 8.74 | 8.76 | 1.39 | 11.5 | 11.5 |
| Length | Adjective suffixes | 2 | 3 | 3.25 | 3 | 0.707 | 4 | 4 |
| | Noun suffixes | 2 | 3 | 3.21 | 3 | 0.704 | 3.75 | 5 |
| | Noun endings | 2 | 2 | 2.94 | 3 | 0.791 | 3 | 5 |

Thus, four noun suffixes, i.e. –ER, –MENT, –NESS, –ISM, for which nonmorphological counterparts were not available, were removed from the relevant analyses (see Table 1). Every participant saw each ending four times (except for the 21 lower-frequency adjective suffixes that appeared eight times). Note that some of our non-morphological originated from Classical languages where those functioned as productive morphemes (e.g., –ME as in "morpheme", "phoneme", "rhizome"; –M as in "rheum"; –LOGUE as in "analogue", "catalogue"; Dee 1984). The design yielded 368 items in total. All stimuli, experimental lists used for presentation, as well as further details on matching across conditions are available on the OSF storage of the project and can be viewed online (https://osf.io/rbxpn/).

Monosyllabic 3-4 letter nonword stems that ended in a consonant were taken from the ARC nonword database (5942 stems; Rastle et al. 2002). These stems were joined with endings. Real words (e.g. lin–EN) as well as homophones (e.g. /dju–tI/) were filtered out. Further, we removed the following: nonwords containing infrequent bigrams (<6 instances per million) and trigrams (<3 instances per million), nonwords that had at least one orthographic neighbour (Coltheart et al. 1977), nonwords with ambiguous endings (e.g. "cli–sy"/"clis–y"), word-like nonwords (e.g. "briber", "bonglike", "lawlist", "thegent"). A manual pronounceability check was not feasible due to a large number of nonwords that were used in this experiment (40112). We minimised the possibility that the presence of "odd" nonwords could influence the results by presenting each participant with a unique combination of stems and endings. Each participant saw a unique experimental list where stems were never repeated.

### *2.1.2 Procedure*

The experiment was implemented online using the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine et al. 2019). The task was to decide if "a letter string looks like a noun or an adjective" by clicking one of the two labelled buttons on the computer screen. It was explained that a noun is the name of something such as a person, place, thing, quality, or idea, and an adjective is a describing word. Real-word examples were given ("time", "people", "way", "year"; "red", "simple", "clever"), and the experiment began with two practice trials that involved real words ("lamp", "colourful") to ensure that participants understood the task. The experiment did not start until the responses on all practice trials were correct. On each experimental trial, participants had eight seconds to respond, otherwise no response was recorded, and the software advanced on to the next trial automatically. For the final seconds of each trial, a countdown clock was displayed in the upper-right corner of the screen. The whole task took, on average, 20 minutes. A progress bar was displayed in the upper left corner of the screen. Trial order was random for each participant. Participants were offered to take three breaks throughout the experiment.

### *2.1.3 Participants*

In order to take part in the study, participants had to be right-handed, British citizens, with no previous history of dyslexia, dyspraxia, ADHD, or any related literacy or language difficulties, raised in a monolingual environment and speaking English as the first language. 109 participants completed the study via Prolific Academic. They were, on average, 24 years old (from 19 to 27 years old); 68 of them were females. One participant indicated that they could also speak French. In terms of education, one participant did not finish high school, 16 finished high school, and 40 finished university. Three participants received professional training, and the rest had a graduate degree.

Average reward per hour was £10.35. Participant read an informed consent form and confirmed that they were willing to take part in the experiment. Since the task was performed online, an extra check was necessary to filter out participants that were not paying attention and/or making little effort to perform well. The main categorisation task did not permit making such judgement, because any response (noun/adjective) was acceptable for any nonword. We opted to use participants' performance on the spelling task as a criterion to filter out poorly performing participants, because in this task, the correct response on each trial was known a priori. Altogether, we excluded three participants whose spellings were further away from the correct spellings (i.e., more than 3 SD away).[1] The distance from the correct spelling was estimated using the Levenshtein distance measure (vwr package in R; Keuleers 2013). For instance, one of the excluded participants produced responses like "youfemism", "apololypse", "bueocrat" for "euphemism", "apocalypse", and "bureaucrat", respectively. Data from 105 participants were retained for analyses.

[1] We also implemented a different, clustering-based algorithm for identifying outlier participants whose behaviour could be different from the rest (Rodriguez and Laio 2014; Borelli et al. 2018). Following this procedure, one participant was excluded in Experiment 1, and none in Experiment 2. All results were replicated. The R script for this analysis is available on the OSF storage for this project.

### *2.1.4 Tasks measuring individual differences*

Vocabulary. Participants completed the Vocabulary sub-scale of the Shipley Institute of Living Scale (Shipley 1940). The vocabulary test consisted of 40 items and required participants to select one word out of four which was most similar to a prompt word in meaning. Response time was unlimited. Vocabulary scores ranged from 14 to 39.

Author recognition. In this test, participants are presented with author names and foils, and are asked to indicate which authors they recognise as real. This test is a reliable predictor of reading skill because author knowledge is thought to be acquired through print exposure (Moore and Gordon 2015; Stanovich and West 1989). The list of 65 existing authors was taken from Acheson et al. (2008). According to an analysis done by Moore and Gordon (2015), the variation in responses that their participants gave to 15 names from this list was minimal and did not have discriminatory power. Therefore, we replaced these 15 names with the names of our choice. These new names were taken from the lists of Pulitzer, Booker, and PEN prizes between 2001 and 2012. We used 65 foil names that were used by Martin-Chang and Gould (2008). Our participants were instructed to avoid guessing as they would be penalised for incorrect responses. The total score was the numerical difference between the number of authors that were identified correctly and the number of authors guessed incorrectly by a participant. This total score ranged from 2 to 49 (out of 65), the mean was 15.

Spelling. Forty words eight letters in length, taken from Burt and Tate (2002), were presented for spelling production. Each word's recording was presented first in isolation, and then a second time in a sentence. The recordings could be replayed for up to 10 times. Participants could type in their spellings after both recordings stopped playing, and they had 15 seconds to do so. A countdown clock was displayed for the last five seconds of each trial. Spelling scores ranged from 0 to 39 (mean was 15).

## 2.2 Analyses

The analyses were performed using generalized linear mixed-effects models (Baayen et al. 2008) as implemented in the lme4 package (Version 1.1-14, Bates et al. 2015) in the statistical software R (Version 3.6.1, R Development Core Team 2018). First, we will present the results of two planned comparisons: (1) adjective suffixes vs noun suffixes; (2) noun suffixes vs noun endings. Two separate linear-mixed models were run to analyse each of these contrasts. Our statistical models included Response as a dependent variable (a binary categorical variable, Adjective coded as 1, or Noun coded as 0), Condition, i.e., ending type (adjective suffix or noun suffix or noun ending, depending on the comparison), as a fixed factor, and random intercepts for subjects and suffixes. Second, we will report the effects of diagnosticity on participants' behaviour. Finally, we will investigate the sources of individual variation in people's sensitivity to these cues.

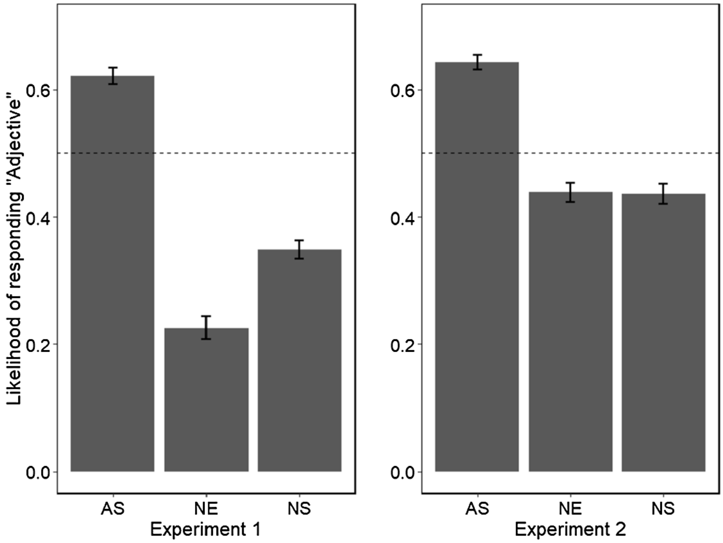

**Fig. 2** Probability of responding "Adjective" (1 on y-axis) or "Noun" (0 on y-axis) for the three types of endings (AS – adjective suffixes, NE – noun endings, NS – noun suffixes). Error bars represent standard errors across participants

### 2.2.1 Item-based variability

Planned comparisons across ending types.[2] The first planned comparison was the contrast between adjective and noun suffixes. As expected, we observed a significant main effect of condition ($z = 5.694$, $p < 0.001$; see Fig. 2) so that more adjective responses were given to nonwords that ended in adjective suffixes compared to noun suffixes. The second planned comparison was the noun suffix versus nonmorphological noun ending contrast, and here as well, we observed a significant difference between the conditions: suffixed nonwords elicited fewer noun responses than nonmorphological nonwords ($z = -3.132$, $p < 0.01$).

In order to understand potential sources of item-based variability, we studied the relationship between ending diagnosticity and participants' responses. Three additional statistical models were implemented separately for each ending type (adjective suffix, noun suffix, noun ending).[3] The models used the continuous measure of diagnosticity as the only fixed predictor (dependent variable as well as random effects were identical to the models described above). The results were as follows.

[2] A combined analysis of all three conditions with "noun suffix" as the reference level and no suffix exclusions replicated the pattern reported here. Namely, nonwords with adjective-biasing suffixes were treated as adjectives more often than those with noun-biasing suffixes ($z = 6.399$, $p < 0.0001$). Noun endings were treated as more noun-like compared to noun suffixes ($z = -3.053$, $p < 0.01$).

[3] A combined analysis of all ending categories showed that the interaction between diagnosticity and ending type was significant ($X^2 = 10.918$, $p < 0.01$) such that there was a positive effect of diagnosticity on the classification of adjective-suffixed nonwords compared to the noun-suffix ($z = -2.262$, $p < 0.05$) and noun-ending ($z = -3.233$, $p < 0.01$) condition, where there was no effect of diagnosticity. There was no difference between the two noun conditions.

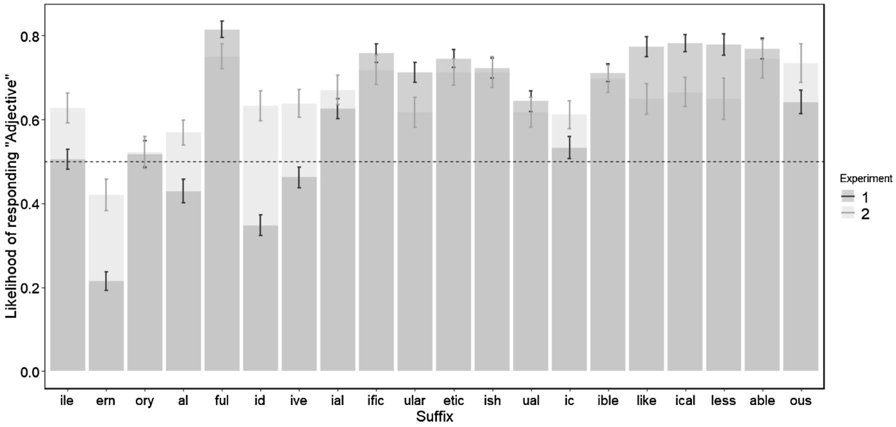

**Fig. 3** Average probability of responding “Adjective” to nonwords that end in adjective suffixes. Suffixes are arranged in the order of increasing diagnosticity. Darker bars correspond to Experiment 1, lighter bars correspond to Experiment 2. As we move from left to right on the x-axis, diagnosticity increases, and the proportion of “Adjective” responses (1 on the y-axes) increases as well

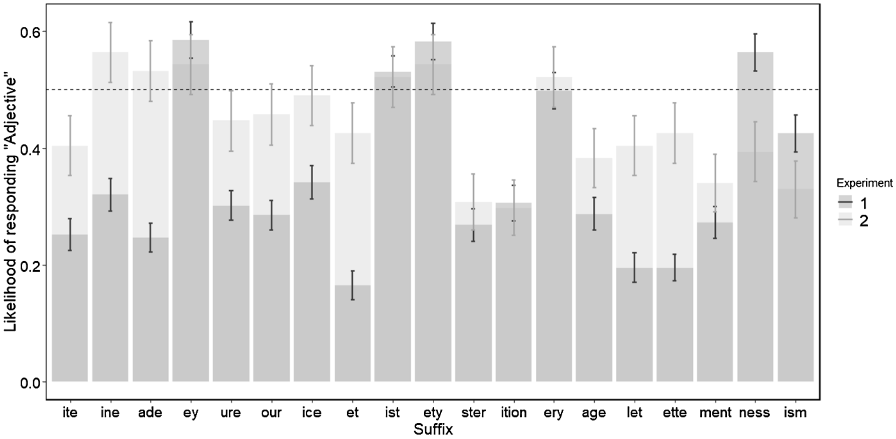

**Fig. 4** Average probability of responding “Adjective” (1) or “Noun” (0) to nonwords that end in noun suffixes. Diagnosticity has no impact on participants’ classification behavior in Experiment 1: as suffix diagnosticity increases, the proportion of adjective (1) responses is constant, and does not decrease, as would be expected, whereas in Experiment 2, suffixes that are higher in diagnosticity elicit more noun responses

Firstly, among adjective suffixes, those with a higher diagnosticity value appeared more adjective-like to our participants ($z = 3.642$, $p < 0.001$), see Fig. 3. The diagnosticity of noun suffixes did not significantly influence responses to nonwords that contained these suffixes, see Fig. 4 ($z = 0.831$, $p = 0.406$). Similarly, we did not observe any impact of diagnosticity of nonmorphological endings on responses to nonwords, see Fig. 5 ($z = -0.983$, $p = 0.326$).

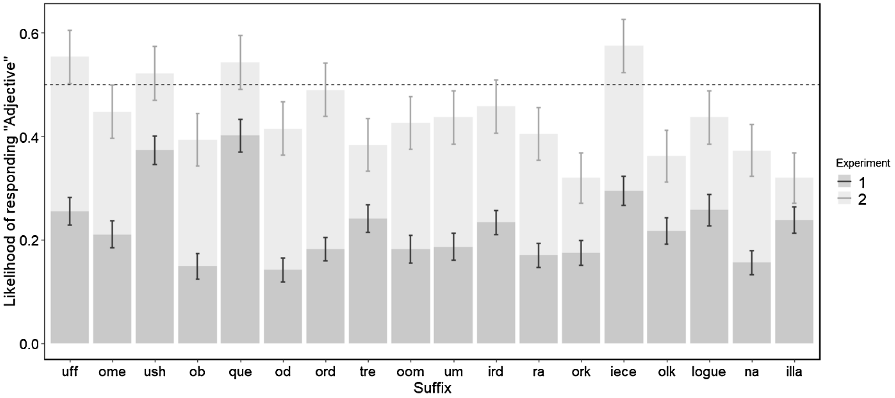

**Fig. 5** Average probability of responding “Adjective” (1) or “Noun” (0) to nonwords that end in noun-biasing nonmorphological patterns. Diagnosticity has no impact on participants' classification behavior: as ending diagnosticity increases, the proportion of adjective responses is constant in Experiment 1, while it decreases in Experiment 2, as expected

**Table 3** Correlation matrix that reflects the relationships between participants' performance on language tasks (ART, vocabulary, and spelling). Spelling scores were sign-transformed for interpretability, so that higher values on all variables reflect better performance

| | Experiment 1 | | | Experiment 2 | | |
|---|---|---|---|---|---|---|
| | Spelling | ART | Vocabulary | Spelling | ART | Vocabulary |
| Spelling | 1 | | | 1 | | |
| ART | 0.50$^{***}$ | 1 | | 0.42$^{***}$ | 1 | |
| Vocabulary | 0.70$^{***}$ | 0.49$^{***}$ | 1 | 0.59$^{***}$ | 0.69$^{***}$ | 1 |

Notes: $^{***}$ corresponds to $p < 0.0001$

### 2.2.2 *Subject-based variability*

In order to address the question of individual differences, we investigated the relationship between participants' performance on background language and literacy measures and their nonword classification performance. Linear mixed models included an interaction between ending type condition and participants' scores on language tasks (three separate models were implemented due to a high correlation between individual characteristics, see Table 3), and potential influences on interpretability of individual effects that this may have (Belsley et al. 2005). The dependent variable as well as the structure of random effects were identical to those in the other analyses reported above. Participants' responses aligned with the predicted lexical category more when participants were better spellers (adjective vs noun suffixes: $z = 19.814$, $p < 0.001$; noun suffixes vs noun endings: $z = -5.189$, $p < 0.001$), had better vocabulary (adjective vs noun suffixes: $z = 16.619$, $p < 0.001$; noun suffixes vs noun endings: $z = -7.630$, $p < 0.001$), or had better author recognition scores (adjective vs noun suffixes: $z = 11.787$, $p < 0.001$; noun suffixes vs noun endings: $z = -3.306$, $p < 0.001$).

## 2.3 Discussion

Experiment 1 replicated earlier findings (Ulicheva et al. 2020). Nonwords with adjective-biasing endings were categorised as adjectives more frequently than nouns. The effects of suffix diagnosticity were graded such that the number of adjective responses co-varied with increasing diagnosticity. We interpret these graded effects as evidence for a statistical learning mechanism that is involved in assimilating these spelling-to-meaning regularities. This conclusion is strengthened when we consider the relationship between participants' performance on language and literacy tests and their sensitivity to category information.

In this experiment, noun endings functioned as stronger cues to category compared to noun suffixes. While these conditions differed in token frequency, as we discovered post-hoc (see *Materials*), we think that it is unlikely that token frequency is responsible for these observed differences. The reason for this view is that noun endings were lower in token frequency than noun suffixes, and as such, they should be *weaker* cues to category. We discuss this finding further in the General Discussion.

One potential flaw of the present experiment is that it involved metalinguistic judgements about lexical category. This is problematic for at least two reasons. Firstly, participants may not have received adequate training as to fully grasp the nuanced distinction between adjectives and nouns. Secondly, the requirement to make metalinguistic judgements could have biased participants to pay attention to lexical category cues. Therefore, in Experiment 2 we opted for an implicit task less prone to these metacognitive influences.

# 3 Experiment 2

The present experiment involved a more implicit version of the category classification task. The replication of Experiment 1 findings in such conditions would constitute stronger evidence in favour of a statistical learning mechanism that is involved in assimilating these cues from the environment.

## 3.1 Methods

### *3.1.1 Materials*

All materials were the same as for Experiment 1, unless indicated otherwise. Every participant saw each noun ending once, and each adjective ending twice, yielding 92 nonwords in total. Nonwords were formed in the same way as for Experiment 1, except that we applied stricter filtering criteria. Specifically, nonwords were used in this experiment if their constituent bigrams occurred at least 10 times per million words (cf. 6 in Experiment 1). In addition, all nonwords (except for those ending in –Y and –Z since there were very few of those) began with legitimate initial trigrams and contained existing quadrigrams. In order to maximise pronounceability, nonwords never embedded frequent (> 6k instances per million words) existing stems (e.g., PAY), doubled consonants, or consonant clusters longer than four letters.

Ninety-two adjective-biasing and 92 noun-biasing sentence frames were created. Sentence frame included a gap denoting placement of the target nonword. Nonwords in the noun context template occupied the syntactic position of a subject or direct/indirect object, usually following an article or an adjective, and so should be interpreted as nouns. Nonwords in adjective contexts appeared after verbs (such as *seem*) and quantifiers (such as *too*) before nouns, which maximised the probability that they would be perceived as adjectives. For example, "He was too ________ for his own good" and "The first ________ was a cathartic experience". Three lists were created that shuffled the pairing of sentence frames with each other, and with suffixes.

### *3.1.2 Procedure*

The task was to decide which of two sentence frames was more appropriate for a target nonword. Two practice trials involving nonwords with an adjective and a noun suffix were constructed using the same criteria as in the main experiment. On each experimental trial, participants had one minute to respond. For the final seconds of each trial, a countdown clock was displayed in the upper-right corner of the screen. The task took about 10 minutes. A progress bar was displayed in the upper left corner of the screen. Trial order was randomised for each participant. Participants were offered a break during the experiment.

### *3.1.3 Participants*

101 participants were tested via Prolific Academic. Four participants were excluded due to technical issues. Inclusion criteria followed Experiment 1. Participants were, on average, 22 years old (from 17 to 26 years old); 59 of them were females. In terms of education, 12 finished high school, and 51 finished university. Three participants received professional training, and the rest had a graduate degree. One participant did not specify their education.

Average reward per hour was £9.82. Participant read an informed consent form and confirmed that they were willing to take part in the experiment. Two participants with abnormally low spelling scores were excluded, following the same procedure as in Experiment 1.

## 3.2 Analyses

A significant main effect of condition ($z = 7.968$, $p < 0.0001$) was observed so that nonwords that ended in adjective suffixes were placed into adjective-biasing sentence frames more often compared to noun suffixes (Fig. 2). There was no difference between the two noun conditions ($z = -.546$, $p = 0.585$). Among adjective suffixes, those with a higher diagnosticity value appeared more adjective-like to our participants ($z = 2.697$, $p < 0.01$; see Fig. 3). Further, noun suffixes with a higher diagnosticity value were more likely to be placed in the noun-biasing context, see Fig. 4 ($z = -2.475$, $p < 0.05$). The same was true for nonmorphological endings, see Fig. 5 ($z = -2.937$, $p < 0.01$).[4]

[4] A combined analysis of all three conditions with "noun suffix" as the reference level and no suffix exclusions replicated this pattern. Nonwords with adjective-biasing suffixes were treated as adjectives more

In terms of subject-based variability, the findings from Experiment 1 were fully replicated. Participants' responses showed greater alignment with the predicted lexical category when they were better spellers (the interaction between spelling and condition was significant, $X^2 = 93.340$, $p < 0.001$, so that better spellers judged nonwords with adjective suffixes to be more adjective-like, $z = 4.356$, $p < 0.001$, and noun-ending nonwords to be more noun-like, $z = 2.954$, $p < 0.01$, with no differences between suffixes and nonmorphological endings). Analogous effects were found for participants with better vocabulary knowledge ($X^2 = 63.372$, $p < 0.001$), and better author recognition scores ($X^2 = 27.037$, $p < 0.001$).

### 3.3 Discussion

Using a task that does not require metalinguistic judgements, Experiment 2 replicated the finding that nonwords with adjective-biasing endings were categorised as adjectives more frequently than nouns. We observed consistent graded effects of diagnosticity across all three ending categories. Further, participants' who performed well on language and literacy tests also exhibited greater sensitivity to category cues. Taken together, our findings strongly suggest that assimilating these spelling-to-meaning regularities involves an implicit, statistical learning mechanisms and is related to reading experience.

## 4 General discussion

In two online crowdsourcing experiments, over 200 participants made overt or covert decisions on whether nonwords resemble nouns or adjectives. Using a large number of endings that varied in how strongly they cued lexical category (i.e., varied in diagnosticity), we replicated earlier findings (Ulicheva et al. 2020; see also Arciuli and Cupples 2003, 2004, 2006; Farmer et al. 2006; Kelly 1992; Arciuli and Monaghan 2009; Kemp et al. 2009). Specifically, nonwords with adjective-biasing endings were categorised as adjectives more frequently than nouns. Further, as suffix diagnosticity increased, the number of category-specific responses increased gradually as well. Following Ulicheva et al. (2020), we interpret these graded effects as evidence for a statistical learning mechanism that is involved in assimilating these spelling-to-meaning regularities. This conclusion is strengthened when we consider the relationship between participants' performance on language and literacy tests (author recognition, spelling, and vocabulary tests) and their sensitivity to meaningful information that is carried by endings. Participants with better language abilities likely have more reading experience, and may have accumulated sufficient lexical and semantic knowledge with which to generalise.

often than those with noun-biasing suffixes ($z = 8.263$, $p < 0.001$). Noun endings were not different from noun suffixes ($z = -0.071$, $p = 0.943$). The interaction between diagnosticity and ending type was significant ($X^2 = 19.810$, $p < 0.001$) such that there was a positive effect of diagnosticity on the classification of adjective-suffixed nonwords compared to the noun-suffixed ($z = -3.858$, $p < 0.001$) and noun-ending ($z = -3.854$, $p < 0.001$) condition. There was no difference between the two noun conditions.

One unexpected result is that we observed a difference in the way morphological and nonmorphological noun endings were categorised in Experiment 1. Specifically, noun endings cued category more strongly than noun suffixes. As discussed in the Introduction, this result appears inconsistent with any existing theory of morphological representation: even theories that assume localist, explicit representations of morphemes would predict an opposite pattern of results, with morphological cues being more salient than nonmorphological ones. Yet, this difference was not present in Experiment 2 where nonwords were presented in sentence context rather than in isolation. Below we propose a potential post-hoc explanation for why the difference between the two types of noun-like nonwords arose in Experiment 1, and not in Experiment 2.

As we alluded to in the Introduction, lexical category is a rudimentary aspect of word meaning, and there may be other meaningful distinctions between morphological and nonmorphological ending types that were not captured by any of our measures. Such distributional differences would be expected to play a role in Experiment 1 where participants were free to draw on any type of information that may be associated with nonwords or their components. On the other hand, these differences may be less pronounced in Experiment 2 where nonwords were presented within a pre-defined syntactic/semantic context thus reducing the need to draw on this type of information, even if it is present within sublexical cues. In the next section, we take the first steps in exploring these distributional cues, and provide some evidence that these might be exploited by participants in both of our experiments.

### 4.1 Distributional differences in meaning and use

Reading experience involves experiencing words in context rather than in isolation (Nation 2017). Words that are experienced in richer linguistic environments enjoy a processing advantage (Hsiao and Nation 2018). On the other hand, words or word parts that behave similarly across contexts may be more related to each other from a cognitive standpoint (Landauer and Dumais 1997). Here we propose that the same may be true for word constituents (cf. morphological family effects; De Jong et al. 2000). Specifically, broader aspects of subword cues' meaning and their use in context may influence how sensitivity to these cues develops with increasing reading experience. In order to explore this idea further, we designed two additional measures that describe these subtle variations in cue meaning and context. We will refer to the first characteristic as "content similarity". It was designed to measure similarity in meaning across all words that carry a given ending. High content similarity values suggest that the ending carries some material meaning and modifies stems in predictable ways. The second measure was "context variability". This captures variability in word use across contexts (see Rinaldi and Marelli, 2020).

In order to operationalise these measures, we applied the distributional semantics methodology (Günther et al. 2019). These techniques are applied to develop data-driven semantic spaces, in which word meanings are represented as vectors that capture lexical co-occurrence patterns from large text corpora. In particular, we adopted the semantic space developed and released by Baroni et al. (2014), induced from a 2.8-billion-word corpus obtained by a concatenation of ukWaC, the English

Wikipedia, and the British National Corpus. The semantic space was trained using the word-embeddings approach proposed by Mikolov et al. (2013), and in particular, the Continuous Bag of Words (CBOW) method. The parameter setting of the adopted space was shown to provide the best performance across a number of tasks (Baroni et al. 2014), largely in line with the psycholinguistic evaluation by Mandera et al. (2017; 5-word co-occurrence window, 400-dimension vectors, negative sampling with k = 10, subsampling with t = 1e-5).

To obtain the first measure of content similarity for a given ending, all words with a given ending that are represented in both CELEX and the semantic space were considered; then cosine similarities for all possible pairs of carrier words were calculated and averaged. For example, 477 words in CELEX end in –OUS (e.g., "fabulous"). Words where the ending is a part of a duplicate stem were removed. In the case of –OUS, six words were removed, such as *subconscious*, because CELEX classifies "conscious" as a stem morpheme and this stem is already counted, i.e., in *conscious*).[5] Further, to allow word pairs with a greater token frequency influence the measure to a greater extent than infrequent pairs, we multiplied each similarity value by the ratio of summed frequency of words comprising this pair to summed frequency of all words containing the ending in question (weighted similarity for –OUS was 0.002, for –IECE was 1.121). For example, suffix –LIKE indicates resemblance (content similarity is 1.165; e.g., "childlike"), while the function of –OUS is less consistent (content similarity is 0.002; cf. "joyous", "nitrous"). This measure is related to the orthography-semantic consistency (OSC) metrics proposed in Marelli et al. (2015), although presenting a critical difference insofar OSC exploits a given free-standing word as pivot for estimating the average semantic similarity between orthographically similar items, whereas content similarity consider all possible word pairs sharing a certain sublexical chunk.

The second measure, context variability, was based on the mean of all vectors corresponding to words with a given ending that are represented in both CELEX and the semantic space (following the approach proposed by Westbury and Hollis 2019). Context variability was then operationalised as the standard deviation of this mean vector. For instance, context variability for –OUS equals 0.057, whereas that for –IECE is 0.070 indicating that words that end in –IECE are used in more variable contexts compared to words ending in –OUS. To illustrate, suffixes –LIKE and –OUS, for instance, are used across a variety of contexts (context variability of 0.061 and 0.057, respectively), whereas the variability of the ending –IRD is higher, 0.076, possibly reflecting the fact that –IRD can occur in verbs ("gird"), nouns ("bird"), and adjectives ("weird", "third").

Nonmorphological endings that we used in our experiment differed from their morphological counterparts (noun suffixes) in content similarity (means were 1.392

[5] The utility of this filtering approach can be demonstrated using an example of an orthographic ending, such as –IECE. Twenty-two words in CELEX end in –IECE. Along with *niece* and *piece*, there are compounds such as *grandniece*, *altarpiece*, *earpiece* etc. The similarity between *niece* and *grandniece* is extremely high, and including such compounds would skew our similarity estimates. By filtering out words with duplicate stems, we are left with only two critical exemplars – *niece* and *piece*. Content similarity of –OUS is 0.164 (the uncorrected value would have been 0.168), of –IECE is 0.138 (the uncorrected value would have been 0.197).

and 0.043, respectively; $t = 3.170$, $p < 0.01$). The difference in context variability was also significant, with noun endings being higher in context variability (mean was 0.061) than noun suffixes (mean was 0.053; $t = 3.786$, $p < 0.001$). We added semantic measures as predictors to the linear-mixed models. Separate models for logarithm-transformed content similarity and context variability were implemented in each experiment, because the two measures were correlated (.62, $p < 0.0001$); glmer(Response∼Diagnosticity*Semantic_Measure + (1|Participant) + (1|Suffix), data). Noun endings that were high on context variability appeared more adjective-like to participants in both experiments (in Experiment 1: $z = 2.442$, $p < 0.05$; in Experiment 2: $z = 1.987$, $p < 0.05$). Noun suffixes were processed in a similar way in Experiment 2: $z = 2.022$, $p < 0.05$. Note that we increased the number of iterations in the statistical models to 25k to let the models converge. In other words, the meanings of cues that are used variably across contexts appear to be somewhat abstract or "diluted"; such words are more readily perceived as adjectives than nouns.

These post-hoc analyses should be interpreted with caution, and our findings should be replicated using more appropriate experimental designs in the future. Firstly, our experiments had not been designed to test the effects of semantic variables, and conditions had not been matched on these. Secondly, the effects of context variability were not found consistently across all categories in both experiments. Nonetheless, these analyses indicate that participants may draw on contextual information when making explicit judgements about lexical category of isolated nonwords, and they do so even when contextual information is readily available to them (as in Experiment 2). We believe that the difference in the design of the two experiments (presentation of nonwords out-of-context as in Experiment 1, or in-context as in Experiment 2) might have influenced the degree to which participants activate and exploit contextual information associated with category cues, resulting in robust differences between noun suffixes and noun endings in Experiment 1, but not in Experiment 2. Our analyses thus suggest that there may be subtle distributional differences in the meanings of cues and their usage in the corpora (see also Gentner 2006; Kemp et al. 2009). Given the requirement to match nonmorphological endings to suffixes on a number of distributional parameters (frequency and diagnosticity), our selected nonmorphological endings were *unusually* high on content similarity and context variability. Consider, for example, the nonmorphological ending –LOGUE that actually used to be a suffix in Latin, meaning "type of discourse" as in "dialogue", "analogue", "catalogue" (Dee 1984). Its content similarity is 8.208, and its context variability is 0.084, while the values of more typical nonmorphological endings that were not used in the experiment due to their low frequency are characterised by lower values (e.g., –ACK with context similarity of 0.036, –IME 0.161). Such endings elicited more category-atypical responses in our experiments.

In the light of research concerning the beneficial effects of consistency for learning (Tamminen et al. 2015; Seidenberg and Gonnerman 2000), we should expect that patterns that ascribe a specific, concrete meaning to stems consistently should be easier to learn than those with a less defined meaning. Furthermore, given evidence on beneficial effects of variability and diversity on learning and word processing (Tamminen et al. 2015; Hsiao and Nation 2018), we should expect patterns that exhibit such variability to be easier to learn. Nonmorphological endings that we used in our experiment thus seem to benefit from both semantic consistency and context variability, and in fact, seem to function like actual morphemes. We therefore believe that these distributional factors likely influence people's responses to orthographic cues. We also believe that connectionist models with a more sophisticated representation of semantic information may be able to capture these regularities in the mapping between spelling and meaning, and thus may be able to account for our results.

Much to our surprise, we found a stark contrast in the way in which words with adjective endings and noun endings were categorised in Experiment 1. Critically, there was no graded effect of noun diagnosticity on responses: all noun-diagnostic cues that we tested were equally strong in eliciting a noun response. However, consistent effects of diagnosticity on all ending types were found in Experiment 2 that used similar materials. The former result might reflect a strategic bias towards perceiving stand-alone words as nouns for the reason that the noun category includes more words than any other category in English, and noun-diagnostic endings are more numerous than endings of any other type (Kemp et al. 2009; Gentner 2006). This result could also reflect the task-related shifts in relative weights that participants place on different types of cues for making category judgements. As we mentioned earlier, in Experiment 1, participants could freely draw on any syntactic and semantic information encoded within the sublexical cues (thus potentially reducing the demand/availability of other types of information, such as diagnosticity). In contrast, rigid sentence frames in Experiment 2 could have constrained the availability or exploitation of sentence-level distributional information.

## 5 Conclusion

In line with Ulicheva et al. (2020), we observed that readers' sensitivity to meaningful cues increased as these indicated lexical category more strongly in written language. We interpret this finding as evidence for a statistical mechanism is involved in the learning of these spelling-to-meaning regularities. We also found that variability across individuals in sensitivity to these cues is related to their performance on language tasks (spelling, vocabulary, and author recognition tests) suggesting that sufficient reading experience is key for developing sensitivity to all types of meaningful cues. In terms of item-based variability, we have identified candidate factors that influenced readers' ability to learn or exploit cues efficiently. These were related to the variability with which these cues are used across contexts.

**Acknowledgements** This work was funded by the Economic and Social Research Council (ESRC) Future Research Leader Fellowship and the Marie Curie Individual Fellowship awarded to A.U. (grant numbers ES/N016440/1 and 747987) and by an ESRC project grant awarded to K.R. (grant number ES/P001874/1). We thank Oxana Grosseck for her contribution to the design of Experiment 2. Authors involved in this work have no conflicts of interest. The OSF storage for this project contains stimuli, data, and additional analyses associated with the work reported in this article (https://osf.io/rbxpn/).

## References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278.

Albrow, K. H. (1972). *The English writing system: Notes towards a description*. London: Longman.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our Midst: An online behavioral experiment builder. *Behavior Research Methods*, 1–20. https://doi.org/10.1101/438242.

Arciuli, J., & Cupples, L. (2003). Effects of stress typicality during speeded grammatical classification. *Language and Speech*, *46*, 353–374. https://doi.org/10.1177/00238309030460040101.

Arciuli, J., & Cupples, L. (2004). The effects of stress typicality during spoken word recognition by native and non-native speakers: Evidence from onset-gating. *Memory and Cognition*, *32*, 21–30. https://doi.org/10.3758/BF03195817.

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *Quarterly Journal of Experimental Psychology*, *59*, 920–948. https://doi.org/10.1080/02724980443000782.

Arciuli, J., & Monaghan, P. (2009). Probabilistic cues to grammatical category in English orthography and their influence during reading. *Scientific Studies of Reading*, *13*(1), 73–93. https://doi.org/10.1080/10888430802633508.

Arciuli, J., McMahon, K., & de Zubicaray, G. (2012). Probabilistic orthographic cues to grammatical category in the brain. *Brain and Language*, *123*(3), 202–210. https://doi.org/10.1016/j.bandl.2012.09.009.

Aronoff, M. (1976). *Word formation in generative grammar* (pp. 1–134). Cambridge: MIT Press.

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

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.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics, Volume 1: Long papers*, 238–247.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Berg, K., & Aronoff, M. (2017). Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy. *Language*, *93*(1), 37–64. https://doi.org/10.1353/lan.2017.0000.

Borelli, E., Crepaldi, D., Porro, C. A., & Cacciari, C. (2018). The psycholinguistic and affective structure of words conveying pain. *PloS one*, *13*(6). https://doi.org/10.6084/m9.figshare.6531308.

Burt, J. S., & Tate, H. (2002). Does a reading lexicon provide orthographic representations for spelling? *Journal of Memory and Language*, *46*(3), 518–543. https://doi.org/10.1006/jmla.2001.2818.

Carney, E. (1994). *A survey of English spelling*. London: Routledge. https://doi.org/10.4324/9780203199916.

Cassani, G., Chuang, Y-Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *46*(4), 621–637. https://doi.org/10.1037/xlm0000747.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

De Jong IV., N. H., Schreuder, R., & Harald Baayen, R. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*(4–5), 329–365. https://doi.org/10.1080/016909600050119625.

Dee, J. (1984). A repertory of English words with classical suffixes: Part II. *The Classical Journal*, *80*(1), 58–62. Retrieved from http://www.jstor.org/stable/3297399.

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, *21*(5), 629–634. https://doi.org/10.1177/0956797610367751.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(32), 12203–12208. https://doi.org/10.1073/pnas.0602173103.

Farmer, T. A., Yan, S., Bicknell, K., & Tanenhaus, M. K. (2015). Form-to-expectation matching effects on first-pass eye movement measures during reading. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 958–976. https://doi.org/10.1037/xhp0000054.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *The Quarterly Journal of Experimental Psychology Section A*, *39*(2), 211–251. https://doi.org/10.1080/14640748708401785.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*(5), 263–279. https://doi.org/10.1017/S0140525X11001841.

Gentner, D. (2006). Why verbs are hard to learn. *Action meets word: How children learn verbs*, 544–564. https://doi.org/10.1093/acprof:oso/9780195170009.003.0022.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. *Perspectives on Psychological Science*, *14*, 1006–1033.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720. https://doi.org/10.1037/0033-295X.111.3.662.

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, *103*, 114–126. https://doi.org/10.1016/j.jml.2018.08.005.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364. https://doi.org/10.1037/0033-295x.99.2.349.

Kemp, N., Nilsson, J., & Arciuli, J. (2009). Noun or verb? Adult readers' sensitivity to spelling cues to grammatical category in word endings. *Reading and Writing*, *22*, 661–685. https://doi.org/10.1007/s11145-008-9140-z.

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package version 0.3.0.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3), 485–515. https://doi.org/10.1037/a0039267.

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1571–1583. https://doi.org/10.1080/17470218.2014.959709.

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284. https://doi.org/10.1111/j.1467-9817.2008.00371.x.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint, arXiv:1301.3781.

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109. https://doi.org/10.3758/s13428-014-0534-3.

Nation, K. (2017). Nurturing a lexical legacy: reading experience is critical for the development of word reading skill. *Science of Learning*, *2*(1), 3. https://doi.org/10.1038/s41539-017-0004-7.

R Development Core Team (2018). *R: a language and environment for statistical computing*. Austria: Vienna. R Foundation, for Statistical Computing. Retrieved from http://www.R-project.org.

Rastle, K. (2019). EPS mid-career prize lecture 2017: writing systems, reading, and language. *Quarterly Journal of Experimental Psychology*, *72*, 677–692. https://doi.org/10.1177/1747021819829696.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: a time-course study. *Language and Cognitive Processes*, *15*(4–5), 507–537. https://doi.org/10.1080/01690960050119689.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: the ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. https://doi.org/10.1080/02724980244000099.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin and Review*, *11*(6), 1090–1098. https://doi.org/10.3758/BF03196742.

Rinaldi, L., & Marelli, M. (2020). The use of number words in natural language obeys Weber's law. *Journal of Experimental Psychology: General*, *149*(7), 1215–1230. https://doi.org/10.1037/xge0000715.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496. https://doi.org/10.1126/science.1242072.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*(9), 353–361. https://doi.org/10.1016/S1364-6613(00)01515-1.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, *9*(2), 371–377. https://doi.org/10.1080/00223980.1940.9917704.

Smith, P. T., Baker, R. G., & Groat, A. (1982). Spelling as a source of information about children's linguistic knowledge. *British Journal of Psychology*, *73*(3), 339–350. https://doi.org/10.1111/j.2044-8295.1982.tb01816.x.

St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, *116*(3), 341–360. https://doi.org/10.1016/j.cognition.2010.05.012.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*(4), 402–433. https://doi.org/10.2307/747605.

Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., Sawi, O., Frost, S. J., & Pugh, K. R. (2019). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading*, *23*(1), 49–63. https://doi.org/10.1080/10888438.2018.1466303.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval or prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647. https://doi.org/10.1016/S0022-5371(75)80051-X.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003.

Treiman, R., Kessler, B., Zevin, J. D., Bick, S., & Davis, M. (2006). Influence of consonantal context on the reading of vowels: evidence from children. *Journal of Experimental Child Psychology*, *93*(1), 1–24. https://doi.org/10.1016/j.jecp.2005.06.008.

Ulicheva, A., Harvey, H., Aronoff, M., & Rastle, K. (2020). Skilled readers' sensitivity to meaningful regularities in English writing. *Cognition*, *195*, 103810. https://doi.org/10.1016/j.cognition.2018.09.013.

Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, *51*(3), 1371–1398.