Behavior Research Methods (2018) 50:1482–1495
https://doi.org/10.3758/s13428-018-1017-8

A database of orthography-semantics consistency (OSC) estimates for 15,017 English words

Marco Marelli[1] · Simona Amenta[2]

Published online: 25 January 2018

Abstract

Orthography–semantics consistency (OSC) is a measure that quantifies the degree of semantic relatedness between a word and its orthographic relatives. OSC is computed as the frequency-weighted average semantic similarity between the meaning of a given word and the meanings of all the words containing that very same orthographic string, as captured by distributional semantic models. We present a resource including optimized estimates of OSC for 15,017 English words. In a series of analyses, we provide a progressive optimization of the OSC variable. We show that computing OSC from word-embeddings models (in place of traditional count models), limiting preprocessing of the corpus used for inducing semantic vectors (in particular, avoiding part-of-speech tagging and lemmatization), and relying on a wider pool of orthographic relatives provide better performance for the measure in a lexical-processing task. We further show that OSC is an important and significant predictor of reaction times in visual word recognition and word naming, one that correlates only weakly with other psycholinguistic variables (e.g., family size, word frequency), indicating that it captures a novel source of variance in lexical access. Finally, some theoretical and methodological implications are discussed of adopting OSC as one of the predictors of reaction times in studies of visual word recognition.

Keywords Orthography–semantics consistency · Form–meaning mapping · Word recognition · Lexical resources · Distributional semantic models

Many factors influence the identification of words presented visually. Behavioral studies, mainly through lexical decision and word naming tasks, singled out a number of properties that affect response times at both the orthographic and semantic levels. The number of orthographic neighbors of a word (e.g., Andrews, 1997; Grainger, 1990), for example, is known to affect its recognition latencies. At the same time, properties associated with word meaning, such as concreteness (e.g., Brysbaert, Warriner, & Kuperman, 2014; Samson & Pillon, 2004), and valence (e.g., Kuperman, Estes, Brysbaert, & Warriner, 2014; Warriner, Kuperman, & Brysbaert, 2013), have shown to affect response latencies in the lexical decision task. Other measures of semantic richness like the number of semantic neighbors, the number of semantic features, or contextual dispersion also have been shown to influence response times in lexical decision and semantic categorization tasks (e.g., Buchanan, Westbury, & Burgess, 2001; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; see also Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012, and Yap, Tan, Pexman, & Hargreaves 2011, for a comprehensive report of semantic richness effects in an array of tasks).

Of course, when investigating visual word recognition, the interface between orthography and semantics is a fundamental issue. On the one hand, orthographic information directly affects visual uptake. On the other hand, semantics is at the core of word comprehension. Early models of visual word identification argued in favor of the activation of purely orthographic representation of words (devoid of meaning) before semantic representations could be accessed. In more recent times, a plurality of studies investigating the processing of polysemous words assigned an important role to feedback semantics at the early stages of word recognition, that is, some aspects of word meaning are activated early on during word recognition and are thus entangled with the orthographic components (e.g., Pecher, 2001; Pexman, Lupker, & Hino, 2002). It has also been shown that during visual word recognition, meanings of orthographic

✉ Marco Marelli
marco.marelli@unimib.it

[1] Department of Psychology, University of Milano-Bicocca, P.zza dell'Ateneo Nuovo 1, 20126 Milano, MI, Italy

[2] Department of Experimental Psychology, Ghent University, Ghent, Belgium

neighbors (e.g., *leopard* and *leotard*) are activated in parallel and this in turn affects semantic categorization (see Rodd, 2004). Similarly, Bowers, Davis, and Hanley (2005) have demonstrated that the meanings of an embedded string and its embedding word are both active (even when there is no morphological or semantic relation between the two of them; e.g., *hat* in *chat*), and this co-activation affects semantic categorization. In conclusion, there is now general agreement around the fact that during word identification the activation of orthographic representations gives way to the activation of semantic properties and these in turn affect the recognition of the word.

The interface between orthography and semantics is a challenging territory for the study of visual word recognition, which could be captured by means of consistency measures. These are not new in the study of word recognition. Phonological–orthographic consistency has received wide attention in the literature (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), indicating that consistent mapping between phonological and orthographic representation facilitates word recognition. Accordingly, on the methodological side it is now easy to automatically estimate the consistency between orthography and phonology (e.g., by relying on computational proposals such as the DRC; Coltheart et al., 2001). However, the interface between orthography and semantics did not receive the same attention (with a few exceptions; e.g., Hino, Miyamura, & Lupker, 2011, comparing Kana and Kanji words in Japanese). This is mostly due to the fact that, although it is easy to trace the boundaries of orthographic units, it is more difficult to capture the semantic dimensions associated with it, especially since semantics is often underspecified if not absent from current reference models of word recognition (e.g., the Bayesian Reader, Norris, 2006; or the original proposal for the Naïve Discriminative Reader, Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; but see the newer implementations by Milin, Divjak, & Baayen, 2017a, and Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017b). Recently, we proposed to investigate the degree of systematicity between orthography and semantics and the impact it has on word recognition by proposing a new, bottom-up measure: The orthography–semantics consistency (OSC; Marelli, Amenta, & Crepaldi, 2015). This measure was developed after observing the result pattern of a large number of morphological masked priming studies, and noticing an overlooked yet consistent side effect: Monomorphemic targets belonging to the "transparent" condition (i.e., the condition in which the related prime is a morphologically complex word that is morphologically and semantically related to the target—e.g., *dealer–deal*) were processed faster than targets in other conditions (e.g., the "opaque" condition, in which the related prime was a pseudo-morphologically complex word that does not entertain any semantic relationship to the target—e.g., *corner–corn*; or the "form" condition, in which the related prime was also monomorphemic and only has a surface relationship with the target—e.g., *scandal–scan*), independently from the fact that they were preceded by a related or unrelated (i.e., random) prime. The explanation of the curious fact (whose validity was backed by a meta-analysis of ten studies in different languages) was indeed found at the interface between orthography and semantics. Let's look at items that were typically included in either the transparent condition—for example, *widow*—or in the opaque condition—for example, *whisk*. Every time the string *widow* is encountered in the lexicon, even as part of other words (*widower*, *widowhood*, *widowed*), it is connected to the concept WIDOW. In this case, we can argue that upon encountering the string *widow*, the probability that it refers to the concept WIDOW is very high, that is, the mapping between the form *widow* and the meaning WIDOW is highly consistent making the string *widow* a reliable cue for the meaning WIDOW. On the contrary, when the string *whisk* is encountered, it may be connected to different meanings. Words such as *whisker*, *whiskered*, *whiskery*, and *whiskey* all embed the string *whisk*, but are not semantically related to the concept WHISK. Hence, we can argue that *whisk* is not a reliable cue for the meaning WHISK, since when it is encountered in the lexicon, it is related to many different meanings and therefore the mapping between the form *whisk* and the meaning WHISK is less consistent. This approach is in line with theoretical proposals originating from learning perspectives (e.g., Baayen et al., 2011; Harm & Seidenberg, 2004). Following this reasoning we can describe orthographic strings as cues that are exploited to reduce uncertainty in the semantic system: The less consistent the form–meaning mapping, the less predictable the status of the semantic system when processing that given form.

To quantify this degree of consistency, and to investigate its impact on word processing, we developed OSC (Marelli et al., 2015). This measure quantifies the relationship between a letter string and the meanings of all the words that share that same sequence in a corpus. We computed OSC exploiting methods borrowed from distributional semantics (Turney & Pantel, 2010), that have shown to be able to provide estimates of semantic association that are sound for behavioral research (e.g., LSA: Landauer & Dumais, 1997; HAL: Lund & Burgess, 1996). The base assumption of this approach is that the meaning of a word can be learned through the way in which it co-occurs with other words in the lexicon. In a distributional semantic model (DSM), word meanings are represented as vectors induced from these lexical co-occurrences. The more two words tend to occur in similar contexts, the more their vectors will be close, and the more their meanings will be considered to be similar. Geometrically, this amounts to measuring the cosine of the angle formed by the two vectors. Capitalizing on this approach, OSC is computed as the frequency-weighted average cosine similarity between the

vector of a word and the vectors of all the words that contain that very same word. OSC is therefore a continuous estimate of the orthography-semantics consistency between a string of letters and the meanings of all the words that contain it.

To verify the ability of OSC to explain response latency on the visual recognition of a wide sample of words, we tested it in a simple lexical decision task, extracting 1821 random words from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012). We showed that OSC scores were significant predictors of reaction times in unprimed lexical decision, over and above family size, word length and frequency effects (Marelli et al., 2015). That is, words with higher scores of OSC (hence, more consistent) are also faster to recognize (see also Jared, Jouravlev, & Joanisse, 2017, for converging evidence).

The relevance of OSC for the visual word recognition literature was further proved in newer studies, in which it was shown that OSC also contributed to explain morphological priming at short SOAs (Amenta, Marelli, & Crepaldi, 2015), and interacted with its phonological counterpart explaining phonological effects in visual words recognition (Amenta, Marelli, & Sulpizio, 2016). These data indicate that, even if the aspect of form–semantics consistency has been rarely taken into account in psycholinguistic studies, it retains indeed a great importance in the study of language processing.

The measure we proposed has noticeable merits, as it is quantitatively and automatically derived, it provides quantitative information that is easy to interpret, and it is mainly atheoretical since it is based on observed quantitative relations between words in a given text corpus. The simplicity of the interpretability of OSC is however countered by the technical expertise and resources necessary to compute it: OSC is easy to use, but requires some technical effort in order to be obtained.

Computing orthography–semantics consistency

OSC was originally computed for a list of 325 monomorphemic target words included in the morphological priming studies considered by Marelli et al. (2015). For each target we first computed a list of "orthographic relatives." We considered, as orthographic relative, each word beginning with the target (e.g., *flux* was a relative of *flu*, but *influence* was not) from a list including the top 30,000 most frequent content words (i.e., adjectives, nouns, verbs, and adverbs) in a 2.8-billion-word corpus (a concatenation of ukWaC, http://wacky.sslmit.unibo.it/, Wikipedia, http://en.wikipedia.org/, and BNC, http://www.natcorp.ox.ac.uk/).

The same corpus (part-of-speech tagged and lemmatized) was also employed to compute the semantic similarity between a target and each of its relatives, defined in geometrical terms through distributional semantics techniques. We focused on word-to-word co-occurrences involving the top 30,000 most frequent content words, collected using a five-word window. Raw counts were reweighted using positive pointwise mutual information (Church & Hanks, 1990), and we reduced matrix dimensions by means of nonnegative matrix factorization (Arora, Ge, & Moitra, 2012), setting the number of dimensions of the reduced space to 350. Mathematically, OSC is computed as the frequency-weighted average semantic similarity between the vector of a word and the vectors of all words that contain that very same word, and can be written as

$$OSC(t) = \frac{\sum_{x=1}^{k} \cos\left(\overrightarrow{t}, \overrightarrow{r_x}\right) * f_{r_x}}{\sum_{x=1}^{k} f_{r_x}}$$

where t is the target word, r_x each of its k orthographic relatives, and f_{rx} the corresponding frequencies. The equation can also be expressed in probabilistic terms:

$$OSC(t) = \sum_{x=1}^{k} p_x * cos\left(\overrightarrow{t}, \overrightarrow{r_x}\right)$$

where p_x is the probability of a given relative in the considered relative set. In these terms, OSC estimates represent the expected semantic similarity between a word and its orthographic relatives (Amenta et al., 2016).

The main scope of the present article is to release a new resource including OSC values for a list of 15,017 words. The measure we describe here represents an improvement with respect to the one used in Marelli et al. (2015) and described above, both in terms of the employed semantic space and the definition of orthographic relatives. In the first analysis we will show how we can achieve an improvement in the performance of OSC by adopting word-embeddings models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), based on neural-network techniques, in place of traditional DSMs (e.g., Turney & Pantel, 2010), based on raw co-occurrence counts. Moreover, we will demonstrate that both part-of-speech tagging and lemmatization of the source corpus are unnecessary preprocessing steps in the computation of OSC values. In the second analysis, we show how we can further improve the OSC performance in predicting response latencies by relaxing the positional constraints in the selection of orthographic relatives, and investigate how the size of the initial pool of potential relatives can influence the measure performance. Finally, we explore the obtained measure and describe its distribution as well as its association with other typical predictors in psycholinguistics.

Analysis 1: The semantic space used

Analysis 1 aims at optimizing the semantic space used to obtain the semantic association estimates that are in turn used for the computation of OSC. The rationale for this analysis is

twofold. First, we turn to word-embeddings models, in place of traditional approaches based on co-occurrence counts, as the former are both better in predicting human responses and more sound from a cognitive point of view (e.g., Mandera, Keuleers, & Brysbaert, 2017). Second, we relax the assumptions at the basis of the original OSC measure, limiting the preprocessing of the corpus used for inducing semantic vectors (in particular, part-of-speech tagging and lemmatization).

The new semantic spaces were trained on a concatenation of BNC, ukWaC, and English Wikipedia, for a total of 2.8 billion tokens, using the freely available word2vec tool (Mikolov et al., 2013). The parameters were set following Baroni, Dinu, and Kruszewski (2014b), who identified as the best performing model, across a number of tasks, the one with the following settings: CBOW (continuous bag of words) method, five-word co-occurrence window, 400-dimension vectors, negative sampling with $k = 10$, and subsampling with $t = 1e{-}5$. Vector representations were induced for each item in the corpus with a frequency of 100 or higher. Three semantic spaces were trained. The first space considered lemma and part-of-speech information in inducing the vectors. As a result, the space encoded different representations for the same word with different grammatical class (e.g., *run* when used as a verb and *run* when used as a noun are associated to different vectors). Moreover, inflectional variants of the same word are associated to a unique vector (the verb forms *speak, speaks, speaking, spoke,* and *spoken* are all associated to the same representation). The second space only considered part of-speech information in inducing the vectors: The space encoded different representations for the same word with different grammatical class (again, *run* when used as a verb and *run* when used as a noun are associated to different vectors), but each inflectional variant is associated to its own representation (each of the verb forms *speak, speaks, speaking, spoke,* and *spoken* is encoded as a separate vector). Finally, the third space did not contemplate any kind of preprocessing information: Each form is encoded as a separate representation, with no differentiation for grammatical class (*run* used as a verb and *run* used as a noun are associated to the same vector). Lemmatization and PoS-tagging were obtained through TreeTagger (Schmid, 1995).

The three semantic spaces were used to compute three OSC measures following the procedure described in the introduction. As the initial relative pool, we considered word forms associated to the top 30,000 content-word lemmas in the corpus. An orthographic relative was defined as a word containing the target item at its orthographic onset (as a result, *corner* was considered an orthographic relative for *corn,* but *scorn* was not). These settings were kept fixed to ensure comparability with the OSC measure we proposed in Marelli et al. (2015). The obtained new OSC measures were labeled as OSC-tagged lemmas (based on the lemmatized and PoS-tagged corpus), OSC-tagged forms (based on the PoS-tagged corpus), and OSC forms (based on the corpus with no preprocessing). The original OSC measure from Marelli et al. (2015) was labeled as OSC-2015.

The quality of the four measures was tested in terms of their ability to explain data variance that is relevant in a cognitive perspective. As a test set, we considered the dataset used in Experiment 3 in the study by Marelli et al. (2015). This consists of 1,818 items (three items of the original set were discarded for technical reasons) with the corresponding lexical decision latencies from the BLP. Table 1 reports the correlation matrix between the four OSC measures in the item set.

Following Marelli et al. (2015), the OSC measures were assessed against a baseline regression model including target frequency, orthographic length, and family size. Frequency was obtained from the RTC Twitter corpus (Herdağdelen, 2013), since Herdağdelen and Marelli (2017) have shown that social media provide the best frequency estimates for explaining lexical decision latencies. Family size estimates were obtained from the morphological annotation of CELEX (Baayen, Piepenbrock, & van Rijn, 1993). Both response times and family size were log-transformed. Frequency was expressed on a Zipf scale (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). We then run a model for each OSC estimate that was included in the analysis along the baseline predictors. Table 2 summarizes the results of the models, reporting for each OSC estimate its effect as well as the variance explained by the corresponding model.

All OSC estimates have significant effects, and provide a significant improvement in the model fit with respect to the baseline (largest p is .0011 at the goodness-of-fit tests comparing each test model with the baseline). The results hence show that the effect of OSC is relatively stable, and different manipulations don't hurt the measure performance, speaking for its robustness. However, a certain variability between the models including different OSC measures is also observed: The results indicate that the performance of the measure can be improved by (a) adopting a better technique to induce semantic vectors (the word-embeddings method by Mikolov et al., 2013) and (b) relaxing the assumptions concerning the source corpus to be used. Indeed, the less preprocessing is applied to the corpus, the better the measure performance, with the highest variance explained observed for the measure

Table 1 Correlation matrix for the four different OSC measures in the item set

	OSC-2015	OSC-Tagged Lemmas	OSC-Tagged Forms	OSC Forms
OSC-2015	1	.75	.71	.63
OSC-tagged lemma	.75	1	.85	.69
OSC-tagged forms	.71	.85	1	.85
OSC forms	.63	.69	.85	1

Table 2 Results of Analysis 1

	Type of Semantic Space	Corpus Preprocessing	OSC Effect		Variance Explained	AIC
Baseline	–	–	–	–	.5383	– 4,823.56
OSC-2015	Traditional DSM	Lemmatization and PoS tagging	$t = 3.29$	$p = .0011$	.5408	– 4,832.37
OSC-tags lemmas	Word embeddings	Lemmatization and PoS tagging	$t = 4.17$	$p = .0001$	.5424	– 4,838.92
OSC-tags forms	Word embeddings	PoS tagging	$t = 4.62$	$p = .0001$	.5434	– 4,842.85
OSC forms	Word embeddings	none	$t = 5.28$	$p = .0001$	.5451	– 4,849.33

The OSC measures are tested against a baseline including word frequency, length, and family size. Associated effects are reported for each OSC estimate, along with the explained variance of the corresponding model.

based on the raw corpus: OSC-forms does not only significantly outperform OSC-2015, based on traditional co-occurrence-count techniques, but also the other measures based on word embeddings (OSC-tags lemmas and OSC-tags forms). It seems that complex preprocessing of the source text data is not needed for using OSC to study word recognition—actually, it even hurts the quality of the measure. In the following analyses, and in the final resource we will hence focus on the OSC measure based on a semantic space that (a) is trained on the raw corpus and (b) adopts the word-embeddings approach.

Having established that corpus preprocessing does not positively contribute to the measure performance, in a follow-up analysis we considered to what extent variations in the parameter space of the distributional model can affect the OSC estimates. Parameters of word-embeddings models have been optimized in previous research efforts, using both online and offline behavioral measures as gold standard (Baroni et al., 2014b; Mandera et al., 2017), and in the present study we indeed adopted a set of parameters associated to good performances in explaining human data. However, it is still interesting to evaluate the robustness of the OSC effect with respect to these parameters. To this purpose, we varied both the size of the co-occurrence window (that defines how large is the context considered during model training in terms of number of words), and the number of vector dimensions (which indicates how many nodes are included in the hidden layer of the network). Moreover, over and above the CBOW approach, we also trained the model using the skipgram system: Whereas in CBOW the estimated weights (i.e., vector dimensions) capture to what extent a target word is reliably predicted by the contexts in which it appears, skipgram weights indicate how well contexts are predicted by the target word. We considered co-occurrence windows of three, five, or seven words, and vectors of either 200 or 400 dimensions, for a total of 12 different semantic spaces (six CBOW models and six skipgram models), and 12 corresponding variants of OSC. These latter were tested against the same dataset and using the same method described above; that is, we evaluated explained variance in (log-transformed) lexical decision latencies for 1,818 items from BLP, using OSC as predictor in a regression model along with Zipf-transformed frequencies, log-transformed family size, and length. The results are reported in Table 3.

As evident from Table 3, changes to the parameter space lead to minimal variations in the measure performance. The OSC effect is very robust to modifications in the training settings of the DSM. Indeed, OSC estimates are very consistent across semantic spaces: The lowest correlation between all possible pairwise comparisons of the 12 OSC variants is $r = .967$ (median correlation is $r = .995$).

Analysis 2: Parameters in the selection procedure for orthographic relatives

In Analysis 2, we move to investigate the effect of the initial pool of orthographic relatives considered, as well as the constraints imposed on the associated search procedure. In principle, these aspects can largely influence the performance of the OSC measure: Search strategies that are not inclusive enough may lead to the exclusion of orthographic relatives of potentially high impact, in turn leading to wrong estimates of the orthographic-semantic consistency associated to a given target. This is particularly evident when considering the positional constraints that we imposed in Marelli et al. (2015), in which we defined as orthographic relatives only words that contained the target at their onset. This criterion excludes not only cases like the *corn–scorn* example reported above, but

Table 3 Effects of the word-embeddings parameters on the OSC measure performance

Co-Occurrence Window	CBOW Models		Skipgram Models	
	200 Dimensions	400 Dimensions	200 Dimensions	400 Dimensions
Three words	.5439	.5448	.5428	.5437
Five words	.5441	.5451	.5426	.5434
Seven words	.5443	.5453	.5426	.5434

Explained variance is reported for regression models including different OSC variants along with word frequency, length, and family size.

also morphologically associated elements such as prefixed words (*replay* for *play*) and compounds (*swordplay* for *play*) that would have an obviously large impact on the OSC estimates. However, the problem is also related to the initial pool of relatives considered: If it is too small, many perfectly acceptable relatives will not be extracted simply because they were not considered as candidates to begin with. On the other hand, we do not want to be too inclusive in the selection of relatives for technical reasons: Vector representations for infrequent words tend to be unreliable because of the sparsity issue in the training data, hence considering such elements in the computation of OSC will only add noise to the estimates.

The potential issues associated to the selection of the relatives can influence the precision of the OSC measure in a negative way. Indeed, very restrictive criteria may lead to most target words being assigned an OSC of 1, that in principle should only happen when a target has only itself as orthographic relative. From a theoretical perspective, OSC = 1 is not necessarily bad. It suggests that the word is a perfect cue for its meaning, as it always appears in orthographic contexts where exactly that meaning is expressed. It is easy to think of examples in which that is really the case—consider long adverbs (e.g., *sympathetically, seamlessly, meaningfully*, etc.): As adverbs, they have no inflected variants, and they rarely act as stem in derived forms, and since they are long they are unlikely to be embedded in other words by sheer chance. However, the interpretation of OSC = 1 is not always so straightforward: Such an outcome may be due to simply erroneous exclusion of relatives because of the restrictive criteria adopted, leading to a wrong estimate of the target property. There is not an easy, automatic way to discriminate between actual case of OSC = 1 and cases in which this latter is the results of uninformative technical issues. Therefore, we aim at reducing such cases by properly setting the selection procedure of the orthographic relatives.

To pursue these objectives, for the present section we evaluated 18 variants of the OSC measure. All these variants were based on the semantic space with the best performance in the previous analysis (that is, the word-embeddings space trained on a non-preprocessed corpus), and differ for the selection procedure for orthographic relatives. In particular, we manipulated the size of the initial pool of relative candidates (ranging from the top 25.000 to the top 65.000 most frequent words in the space, with an interval of 5,000 words) and the constraint concerning the position of the target word within the potential relatives (word onset vs. no constraints). OSC estimates were obtained for all the content words (nouns, verbs, adjectives, and adverbs) found in both the BLP and our semantic space.

The obtained measures were first tested in terms of their ability to predict lexical decision latencies in a new test set. In fact, the set used in the previous analysis included items from Marelli et al. (2015) that, although randomly extracted, were based on an initial word sample related to the derived words used in morphological priming experiments. As a result, they could have been biased toward the onset-locked measures, making them an unfit test set for the present evaluation. Therefore, in this analysis we extracted, as test items, all words in BLP for which OSC was different from 1 for all the 18 variants considered. The resulting set consisted of 3,065 words. As in previous analyses, the OSC measures were assessed against a baseline regression model including target frequency, orthographic length, and family size. Frequency norms were obtained from the RTC Twitter corpus (Herdağdelen, 2013), family size estimates were obtained from CELEX (Baayen et al., 1993), and lexical-decision latencies were extracted from the BLP. These three variables were log-transformed (frequency on a Zipf scale; Van Heuven et al., 2014). We then tested a model for each of the OSC variants computed, by including this latter measure alongside the baseline predictors.

Figure 1 summarizes the results of the analysis, reporting for each OSC variant the explained variance of the corresponding model. Additional information is reported in Table 4. There is not much variability in the measure performance with respect to the size of the initial relative pool: The observed explained variance is relatively stable across the different sizes considered. However, the results indicate a negative impact of the positional constraint: Although for both conditions an improvement with respect to the baseline is observed, said improvement is more marked for the case in which the relative selection is not limited to candidates that share the orthographic onset with the target (see also Bowers et al., 2005). In line with Analysis 1, these results (i) speak for the robustness of the measures, whose performance is essentially unaffected by the pool size considered, and (ii) indicate

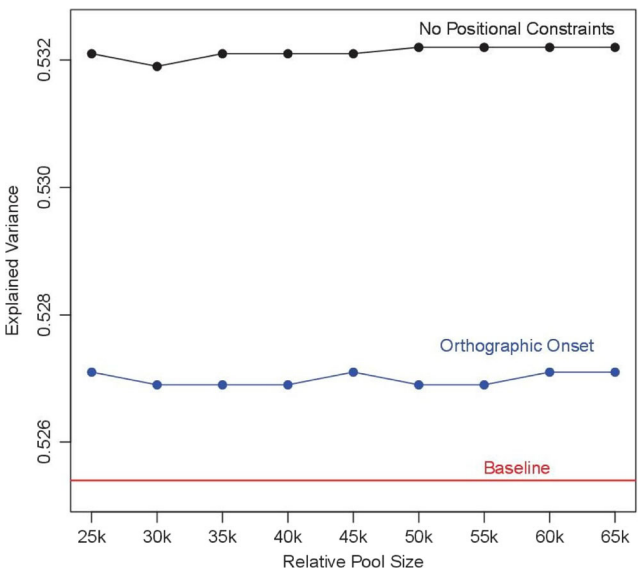

Fig. 1 Impacts of the size of the initial relative pool and the onset positional constraint on the measure's performance.

Table 4 Performance of different OSC variants, showing the impacts of the size of the initial relative pool and of the onset positional constraint

Initial Pool Size of Relative Candidates	Onset-Locked Relative Selection	OSC Effect		Variance Explained	AIC
Baseline	Baseline	–	–	.5254	– 6,452.43
25.000	TRUE	$t = 3.38$	$p = .0007$	.5271	– 6,461.84
65.000	TRUE	$t = 3.33$	$p = .0009$	.5269	– 6,461.50
25.000	FALSE	$t = 6.69$	$p = .0001$	.5321	– 6,494.99
65.000	FALSE	$t = 6.75$	$p = .0001$	.5322	– 6,495.73

The OSC measures are tested against a baseline including word frequency, length, and family size. Associated effects are reported for each OSC variant, along with the explained variance of the corresponding model.

that we can further relax our assumptions concerning the computation of OSC, namely by dropping the positional constraints in the relative selection.

However, this does not indicate that a larger candidate pool for orthographic relatives is not useful at all. For the sake of comparability between measures, we have limited our comparison to a set of words for which *all* variants had OSC different from 1. However, as we argued above, it is the case that the present manipulations can influence the probability of producing erroneous estimates of OSC = 1, an outcome that we would like to limit as much as possible. Hence, we evaluated to what extent the probability of predicting OSC = 1 changes in relation to the initial relative pool size and the positional constraints in relative selection. The results of the analysis are represented in Fig. 2. These data are based on the full set of 15,017 words for which we have obtained OSC estimates.

As expected, the parameters have an impact on the estimated proportion of OSC = 1 items. In line with the previous analysis, positional constraints have an undesired impact, raising this proportion in comparison to the condition without such constraints. Moreover, the initial pool size has also an effect: The larger the pool of initial candidates, the lower the proportion of OSC = 1 items. This outcome has to be expected: Both the lack of positional constraints and a larger candidate pool makes orthographic candidates more likely to be found, in turn leading to more well-distributed (and arguably more reliable) OSC estimates. On the basis of these results, we opt for focusing on the measure without positional constraints and based on the largest relative pool for the released resource.

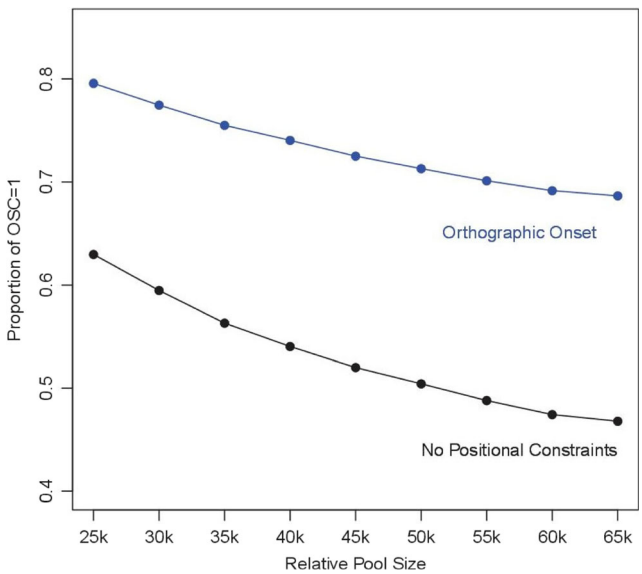

Fig. 2 Impacts of the size of the initial relative pool and the onset positional constraint on the proportion of words for which OSC equals 1.

However, until now we have defined orthographic relatives considering the whole string constituting the target word: An orthographic relative is a word that fully embeds the target (relatives of *part* are, for example, *partial*, *apart*, *partner* . . .). This stems naturally from our initial objective of explaining human performance in word-processing experiments, in which words are typically presented as orthographic strings in isolation. The operationalization of the orthographic relatives (and consequently of OSC) moves from this practical, atheoretical observation, and hence uses the very strings presented to participants as probes to extract the orthographic cohorts related to each target. Although some works in the psycholinguistic literature have followed a similar approach (e.g., Bowers et al., 2005), other studies on the effect of a word's orthographic neighborhood have considered more nuanced operationalizations of this aspect. An ideal example in this respect is the work by Yarkoni, Balota, and Yap (2008), in which orthographic relatives are not defined in binary terms, but rather as a function of their orthographic distance from the target: A word can be more or less "neighborly" with respect to a given target depending on the Levenshtein distance (LD; Levenshtein, 1966) between the two strings. According to this metric, the distance between two words is the minimum number of substitution, insertion, or deletion operations required to turn one word into the other (e.g., the LD between *cast* and *cost* is 1; the LD between *cast* and *casting* is 3; and the LD between *cast* and *costing* is 4). Since the LD has been showed to be an ideal metric to characterize orthographic neighborhoods in word recognition studies (Yarkoni et al., 2008), one may wonder whether using such metric to define orthographic relatives for the computation of OSC could further improve the measure performance.

To assess this possibility, we used the LD between a target and each other word in the lexicon as an index to select orthographic relatives, in turn obtaining six novel variants of OSC. Three of these variants (OSC-top10, OSC-top20, OSC-top30) were based on relative sets defined in terms of top orthographic neighbors: OSC was computed as the (frequency-weighted) average semantic similarity between a target and its *n* orthographically closest word, where *n* could be 10, 20, or 30. In the other three variants (OSC-LD1, OSC-LD2, OSC-LD3), relatives were extracted on the basis of a previously defined maximum distance: OSC was computed as the (frequency-weighted) average semantic similarity between a target and all the words with a maximum LD of *n* from it, where *n* could be 1, 2, or 3; in other terms, in these latter variants relatives where defined as words that could be obtained by deleting, adding, or substituting 1/2/3 letters from the target. Semantic estimates used to compute these OSC variants were based on the same word-embeddings space employed in the previous evaluations, and we considered as initial pool of relative candidates the same set of 65.000 words presented above.

The performance of the novel six OSC variants was assessed using the same procedure described in the previous evaluation: We employed a dataset of 3,065 items from the BLP, and we computed the variance explained in human latencies by a regression model including OSC along with Zipf-transformed frequency, log family size, and length. Table 5 reports the results of this evaluation, comparing the performance of each of the novel six OSC variants with the best-performing OSC measure from the previous analysis.

The analysis showed that, when OSC is computed on the basis of LD-defined relatives, its ability to explain variance in human behavior is lower than when orthographic relatives are selected through the substring approach described in Marelli et al. (2015). The approach based on LD still looks promising, with significant effects when the relatives are defined in terms of a maximum LD from the target, but it is largely outperformed by the original characterization of OSC. This suggests that, in this kind of word-processing experiment, semantic information may be mainly accessed through the whole string constituting a word, with subword orthographic units playing a more limited role in informing meaning. Of course, this does not imply that LD-based measures, such as OLD20 (Yarkoni et al., 2008), are not viable options to characterize neighborhoods on a purely orthographic level; rather, it suggests that, when the purpose is capturing orthographic-semantic relations, focusing on orthographic chunks leads to more informative estimates.

Description of the resource and evaluation on megastudies

We release a dataset of OSC values for the content words included in the BLP (Keuleers et al., 2012). The final set includes 15,017 items. The reported OSC measure is based on the best-performing parameter settings, as estimated by the analyses described in the previous sections. We excluded grammatical words from the present item set since it is typically difficult to obtain high-quality semantic vectors for such element (Baroni, Bernardi & Zamparelli, 2014a): In these cases, contexts are typically not particularly informative, since grammatical words tend to frequently co-occur with all the other words in the corpus (i.e., they have low predictive power). The uninformative vectors that are usually obtained for grammatical words, paired with their extremely high frequency, would lead to low-quality OSC estimates. We also opted for not including in the database words with very low frequency, since these items may present an issue that is opposite to the one described for grammatical words: DSMs cannot induce reliable vector representations for rare words, because of the scarcity of training data (Turian, Ratinov, & Bengio, 2010). Finally since, in principle, every word is an orthographic relative of itself, all the items of the database are also included in the pool of 65,000 words from which orthographic relatives were selected. The exact meaning similarity of a word with itself is a crucial piece of semantic information in

Table 5 Performance of different OSC variants when using LD to select orthographic relatives

	Type of Relatives	OSC Effect		Variance Explained	AIC
Baseline	–	–	–	.5254	– 6,452.43
OSC	Words that embed the target	$t = 6.75$	$p = .0001$	.5322	– 6,595.73
OSC-top10	Top 10 LD-defined neighbors	$t = 0.27$	$p = .7821$	.5254	– 6,450.50
OSC-top20	Top 20 LD-defined neighbors	$t = 1.75$	$p = .0805$	.5257	– 6,453.49
OSC-top30	Top 30 LD-defined neighbors	$t = 1.41$	$p = .1601$	.5256	– 6,452.40
OSC-LD1	Words with maximum LD = 1	$t = 2.19$	$p = .0288$	.5259	– 6,455.21
OSC-LD2	Words with maximum LD = 2	$t = 2.03$	$p = .0421$	.5259	– 6,454.57
OSC-LD3	Words with maximum LD = 3	$t = 1.36$	$p = .1751$	.5255	– 6,452.27

OSC measures are tested against a baseline including Zipf-transformed word frequency, length, and log family size. Associated effects are reported for each OSC estimate, along with the explained variance of the corresponding model.

the computation of OSC since it is the only relation, within the activated orthographic cohort, that perfectly captures the meaning associated to the target. Words that were not part of this relative pool are hence not included in the released resource because their corresponding OSC estimates could not be properly computed.

In the present section we first investigate the proportion of items for which OSC equals 1 in the final dataset. As we discussed, it is difficult to establish a priori whether these are erroneous estimates related to technical limitations, or proper cases in which a word has no orthographic relatives to be found. The manipulations described in the previous section has lowered the number of OSC = 1 items, arguably correcting a number of erroneous estimates. However, a large number of such items can still be observed (about 46% of the words). Table 6 suggests why this may be the case (the considered variables were obtained as described in the previous sections).

Items with OSC = 1 are less frequent, are longer, and have smaller family sizes than items whose OSC is not 1. This pattern of results is sensible for cases for which OSC = 1 is a reliable estimate. In fact, smaller family sizes make it less likely for a word to be related to derived forms and/or compounds that would be included in its relative selection; longer words are less likely to be embedded in other words by sheer chance, hence further diminishing the probability of having orthographic relatives; and less frequent words would have inflectional variants (if any) that are even less frequent, hence they won't be found in the lexicon of the semantic space. To summarize, the results suggest that OSC = 1 items could be, to a large extent, genuine cases.

A large number of OSC = 1 items makes sense if we consider the lexicon a set of signs shared by a number of speakers. To ensure communication, we would not expect such a system to be extremely arbitrary.[1] That is, we may expect that most words will be good cues for their meanings (e.g., Dingemanse et al., 2015; Monaghan & Christiansen, 2006). Indeed, this is confirmed by the distribution of OSC (when it is different from 1) in the whole dataset (Fig. 3): Even when not considering the OSC = 1 cases, the variable is quite skewed, with many word having OSC higher than .80. The distribution of items with OSC equals to 1 vis-à-vis different from 1 could be related to hapax legomena. This cannot be tested directly with the present approach since, for technical reasons, high-quality semantic vectors for rare words are very difficult to obtain (e.g., Turian et al., 2010), and the OSC computation is based on the availability of reliable semantic representations. We may nevertheless conjecture that hapax legomena would be quite in line with our description of OSC = 1 items, to the very least when they represent relatively novel lexical entries. In fact, these elements are often the consequence of morphological productivity (affixation or compounding processes), thus they tend to be orthographically long words. For this reason, they are unlikely to be subject to morphological operations themselves, and hence to be included in complex forms. Similarly, being so rare they are unlikely to be observed also in inflected forms. As a result, hapax legomena will rarely have orthographic relatives over and above themselves, hence resulting in a OSC of 1. Indeed, these considerations about hapax legomena fit well the data pattern described in Table 6: OSC = 1 items tend to have lower family size, higher orthographic length, and lower frequency.

Table 6 Comparison of items for which OSC equals 1 vis-à-vis items for which OSC is smaller than 1

	Items in which OSC = 1	Items in which OSC < 1	*t*	*p*
N	7,027	7,990	–	–
Frequency	1.76 ± 0.016	2.87 ± 0.017	47.56	.0001
Family size	0.38 ± 0.006	1.26 ± 0.011	74.61	.0001
Length	6.94 ± 0.017	5.49 ± 0.017	60.03	.0001

Finally, we test OSC on latencies from megastudies. We extracted response latencies from the BLP and focused on items for which OSC is different from 1 (7,990 data points). Similarly to previous analyses, we also included in the regression model item length, Zipf-transformed frequency, and log-transformed family size. OSC has a negligible correlation with the latter measure (r = .09), and a moderate correlation with the former two (r = .43 and .27, respectively). The effect of OSC on response latencies is however significant (t = 4.51, p

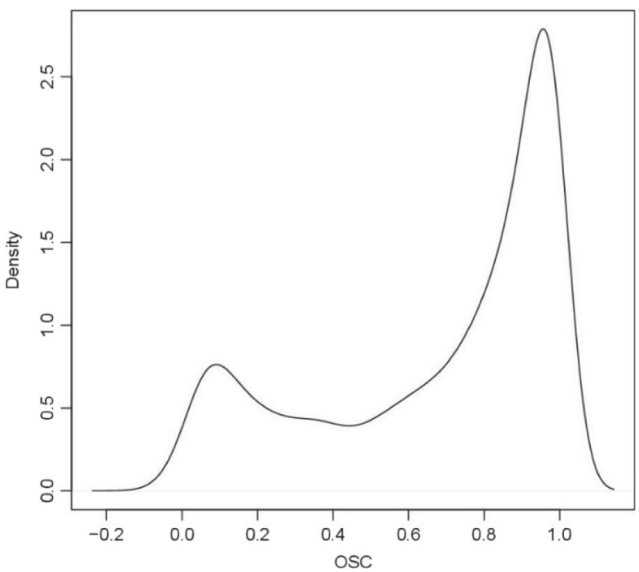

Fig. 3 OSC density distribution.

[1] With this term we do not make reference to the fact that words are arbitrary symbols for concepts; rather, we point out that, at a lexical level, words that have associated orthographic forms often share associated meanings. The term is thus intended to be interpreted exclusively in a "distributional" sense, referring to the distribution of orthographic forms in the lexicon (see Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015, for further discussion).

= .0001) and holds also when removing from the whole dataset items with average accuracy lower than .66 (following Brysbaert & New, 2009; t = 7.74, p = .0001). In both cases, the higher OSC, the faster the response times, in line with the results reported in Marelli et al. (2015).

However, the covariates we considered do not exclude the possibility of a semantic confounding in the OSC effect. In particular, since the measure is based on meaning association estimates from DSMs, it is in principle possible that the effect of OSC is simply explained by the semantic dimensions captured by the vector model, rather than a combination of orthographic and semantic features. To exclude this hypothesis, we adopted the methodology proposed by Hollis and Westbury (2016), who showed that semantic metrics obtained by applying principal-component analysis on a DSM are significant predictors of lexical decision latencies. Following Hollis and Westbury's (2016) procedure, we applied principal component analysis on the very same semantic space we used to compute OSC, after having mean-centered each dimension and scaled it to have unit variance. The procedure resulted in 356 principal components (PCs) accounting for 95% of the variance between vector dimensions. These 356 PCs were included as predictors in a regression model on the BLP response times along with Zipf-transformed frequency, log-transformed family size, orthographic length, and OSC. The analysis was run on the same set of 7,990 words described above. Of the 356 PC predictors, 34 were found to have a significant effect at p = .10. However, their inclusion does not affect the impact of OSC, whose effect holds in the new analysis (t = 4.27, p = .0001). The results of this test indicate that the OSC effect cannot be reduced to the impact of semantic features directly extracted from the source DSM. The evidence supports the OSC characterization as a hybrid measure capturing the mapping between orthographic and semantic information.

Testing OSC in predicting the response times from the BLP is straightforward, and consistent with the source data used to compute the measure (mostly based on British English). However, it is known that such source data can have an impact on measure performances, even when considering relatively similar language variants such as British and American English (see Keuleers et al., 2012, and Herdağdelen & Marelli, 2017). Therefore, it is important to establish that our OSC estimates are also significantly associated with data from American English. We thus turn our attention to the ELP (Balota et al., 2007), that collects response latencies in lexical decision from American speakers. Moreover, the ELP data also offer the opportunity to test the measure performance on different experimental tasks, since they also report latencies for word naming.

For the analysis on the ELP data, we considered all the items that are also included in our OSC set. Similarly to the procedure above, we focused only on the items with OSC different from 1 (7,108 data points), and we included as predictors also Zipf-transformed frequency, log-transformed family size, and length. When considering lexical decision latencies, the effect of OSC is significant (t = 4.97, p = .0001) and holds when considering only items with average accuracy higher than .66 (t = 5.59, p = .0001). A similar pattern of results is observed for response times in word naming, with again significant effects for both item sets (t = 3.17, p = .0016; t = 2.74, p = .0061). The results on the ELP speak again for the robustness of the OSC measure we release, for which we observed solid effects across language variants (British vs. American English) and experimental tasks (lexical decision vs. word naming).

As a further test of the measure robustness, we evaluated the OSC effect when controlling for frequency estimates from different sources. In fact, since frequency plays an important role in the computation of OSC, it is possible that OSC accounts for behavioral variance above and beyond word frequency because it has been calculated over a different language distribution than the frequency measure employed in the previous analyses. We thus considered frequency estimates from the corpus used to compute OSC (a concatenation of BNC, ukWaC, and English Wikipedia), along with frequency norms based on television subtitles (SUBTLEX-UK: Van Heuven et al., 2014; and SUBTLEX-US: Brysbaert & New, 2009; Brysbaert, New, & Keuleers, 2012). The OSC interplay with these frequency measures was evaluated in the three datasets considered above, including items with OSC different from 1. For each of these datasets, Table 7 reports the correlation between OSC and log-transformed frequency (in terms of Kendall's tau and Pearson's r) and the OSC effects in a regression model when including, as covariate, the respective log-transformed frequency measure, along with orthographic length and (log-transformed) family size.

Table 7 shows that the OSC impact on behavioral data is relatively stable across different frequency covariates. The effects are smaller when frequency norms are obtained from the same source used to calculate OSC, which is to be expected since lexical frequencies are included in the OSC computation. However, importantly, the effect of the measure holds also in these cases.

A further influence of lexical frequency on OSC performance could in principle emerge from the orthographic relatives included in the OSC computation, since the semantic impact of these latter is weighted for their frequency. Indeed, one may wonder to what extent OSC may simply reflect the inhibitory effect of high-frequency relatives. However, this is highly unlikely. Certainly the impact of the different relatives of a given word will depend on their frequency, but it won't be necessarily inhibitory: Actually, the way OSC is formalized ensures that relatives can have either an inhibitory (i.e., leading to lower OSC values) or facilitatory role (i.e., leading to higher OSC values), depending on their semantic association

Table 7 Interplay between OSC and frequency norms from different corpora

	BLP			ELP			
	Kendall's Tau	Pearson's *r*	OSC Effect Lexical Decision	Kendall's Tau	Pearson's *r*	OSC Effect Lexical Decision	OSC Effect Word Naming
RTC (Twitter)	.171	.269	$t = 4.51$	.113	.171	$t = 4.97$	$t = 3.17$
BNC + UkWac + Wikipedia	.276	.369	$t = 2.49$	.239	.319	$t = 2.45$	$t = 2.67$
SUBTLEX-US	.169	.219	$t = 6.55$	.121	.143	$t = 5.85$	$t = 1.84$
SUBTLEX-UK	.157	.196	$t = 6.93$	.128	.148	$t = 5.61$	$t = 2.09$

The table reports (1) correlations (in terms of Kendall's tau and Pearson's *r*) between OSC and frequency and (2) the effect of OSC when the corresponding frequency estimates are included as a covariate in a regression model against behavioral data.

with the target. In line with this theoretical consideration, we fail to observe any correlation between the average frequency of the orthographic relatives and the corresponding OSC estimates ($r = .004$); moreover, the OSC effects reported in these section analyses hold when including the average frequency of the relatives as covariate in the regression models.

Conclusions

In the present work, we present the orthography–semantics consistency (OSC) measure as a valid and effective predictor of response latencies in visual word recognition, tested on a large number of words in lexical decision and word naming paradigms in American and British English. We provide a complete resource containing OSC values for 15,017 English content words, so that the measure can be readily available and easily accessible by the entire psycholinguistic community. Along the OSC norms we release the corresponding log-transformed frequency values (as extracted from our source corpus) for experimental-control purposes. The complete database can be downloaded from www.marcomarelli.net/resources/osc.

The OSC measures provided here is the most updated and best-performing version to present. In a series of analyses, we have in fact shown that previous constraints imposed on the measure led to a worse performance in predicting response times, in comparison to the new and improved version. With respect to the considered parameters, we propose here a simplified version of OSC that does not rely on part-of-speech tagging and lemmatization and is free from positional constraints in the selection of orthographic relatives. Moreover, we extended the search pool for orthographic relatives to the top 65,000 most frequent content words in the lexicon. This version proved to explain the most variance in lexical decision (and proved to be a significant predictor also in word naming). It is worth noting, however, that although the measure that we release is the best option at the moment, all the versions tested in this article were valid predictors of significant portions of variance in lexical decision. OSC therefore proved to be a robust measure, not particularly sensitive to changes in the associated parameters.

From a theoretical perspective, the results presented in this study are informative of the role of meaning in word recognition. As a measure, OSC has a strong semantic nature: It quantifies the mapping between orthographic strings and their associate meanings. Its impact in visual word recognition is further proof of the crucial role played by semantics in lexical access, even in tasks whose nature is not explicitly semantic (e.g., lexical decision). Our results are also consistent with Bowers et al. (2005) data on subset–superset activation. In their seminal work, Bowers and colleagues showed that subsets (e.g., *hat*) are activated independently from their position within the superset (e.g., *hat*ch or c*hat*), and that the semantics of the subset interferes with the processing of the superset. Their results have implications for both orthographic-coding theories and theories on semantic activation. Our results provide converging evidence for both these stances. Relative to orthographic coding, it should be noted that our best performing formalization of OSC includes the selection of orthographic relatives that embed the full target word in either initial, middle, or final position. In other words, the semantics of orthographic relatives contributes to the definition of the target OSC independently of the position of the embedded string (i.e., the target word). As Bowers and colleagues suggested, this poses an issue for models of orthographic encoding based on serial activation (e.g., Coltheart et al., 2001), and support alternative proposals based on parallel coding (e.g., Davis, 1999; Davis & Bowers, 2004). Concerning semantic activation, OSC has been described in this article as a measure of similarity between the meaning of a target word and the meanings of its orthographic relatives. In this sense, we can redefine OSC as an index of orthographically informed semantic activation, as it indicates that, when we read a word, not only the orthographic form of its relatives is activated, but also their semantic representations. Hence, our results support the conclusions of Bowers and colleagues relative to the cascaded activation of semantic representation from orthographic input. It is worth noting, however, that differently from Bowers et al. our task did not ask for an explicit semantic judgment, providing further evidence that

word meaning is active in word recognition also when the experimental task does not invoke it directly. In this respect, the effect of OSC on word naming data is particularly interesting. The effect of semantics in word naming has produced divergent results, often depending on the way that semantics was operationalized (see, e.g., Bates, Burani, d'Amico, & Barca, 2001, vs. Buchanan et al., 2001). In our study, semantics was operationalized in distributed terms, in a way related to the one intended by Buchanan et al. (2001), which used HAL to quantify semantic neighborhood size. Indeed, our results are coherent with their work, as in both our and their analyses semantic effects are observed. However, even if founded on similar distributional underpinnings, OSC provides a different semantic characterization. On the one hand, in the present article the measure we computed is based on word embeddings, an approach that has proven to outperform traditional distributional models in a number of different tasks (Baroni et al., 2014b; Mandera et al., 2017), and produces more nuanced and cognitively plausible representations (Keith, Westbury, & Goldman, 2015; Mandera et al., 2017). On the other hand, OSC captures semantic information that is tightly entangled with the word orthography and has an effect on lexical access that is independent from the one associated with the sheer semantic neighborhood (Amenta et al., 2015). As such, OSC describes a hybrid component that provides consistent but novel evidence concerning the role of semantics in word naming. The composite nature of the phenomenon captured through OSC is theoretically in line with word learning models that builds on reliable patterns between forms and meanings (e.g., Baayen et al., 2011; Harm & Seidenberg, 2004). Indeed, it is conceivable that the OSC effect could be a by-product of such learning systems. For example, recent developments of the NDL architecture (e.g., Milin et al., 2017), that integrate semantics as associations between higher-level units (lexomes), could be able to provide a computational explanation of the OSC impact in word recognition.

From a methodological perspective, the resource presented here ideally complements previous efforts to quantify consistency in language systems, typically focused on the mapping between orthography and phonology (e.g., Balota et al., 2004), and constitutes an important source of data to be considered by researchers interested in word recognition. In fact, the results by Marelli et al. (2015) indicate how not accounting for the OSC contribution could lead to serious confounding, and be the source of unexpected (and apparently inexplicable) side effects. In the present article, we further provide evidence for the impact of OSC in large datasets, with effects that are reliable across different tasks and language variants. Moreover, we show that OSC correlates only weakly with family size, capturing a different, largely unexplored component of the word recognition process. On par with this measure, it is recommended to always address the role of OSC or, at the very least, to include it as a covariate in studies of visual word recognition.

The present focus of the proposed database on a relatively limited set of items (i.e., a set that excludes most function and rare words) is mostly related to present technical limitations of DSMs, rather than of the OSC measure itself. This particularly pertains to low-frequency words. Certainly, vectors for rare words can be directly obtained through an incremental procedure such as the CBOW one, but such representations tend to be particularly noisy due to data scarcity (Turian et al., 2010). This is not necessarily a drawback if DSMs are taken as models of how humans can extract meaning from language usage (Mandera et al., 2017): Similarly to a distributional model, human speakers will not have a precise idea of the meaning of a word from observing it in context just once or twice. On the other hand, this limitation may be a problem if one's scope is to obtain reliable representations for practical purposes (e.g., developing a resource, like in the present study). Solutions to this problem are being proposed, often focused on exploiting meanings of sublexical strings to induce representations for low-frequency words (e.g., Bojanowski, Grave, Joulin, & Mikolov, 2016; Lazaridou, Marelli, Zamparelli, & Baroni, 2013) and even novel words that never appear in the training corpus (Marelli & Baroni, 2015). Other methods rely on enriching the new semantic vector through multimodal data, based on the lexical context in which the rare word is found (Lazaridou, Marelli, & Baroni, 2017). As for grammatical words, promising approaches have been advanced in the domain of compositional distributional semantics (Baroni et al., 2014a; Bernardi, Dinu, Marelli, & Baroni, 2013). These ongoing developments may provide the possibility of inducing higher quality vectors for such problematic elements in the next future, and hence allow databases of measures relying on DSMs (such as OSC) to be further developed.

Author note Authors contributions are as follows: M.M. and S.A. designed the study; M.M. developed the models and ran the statistical analyses; M.M. and S.A. drafted the article. This study was supported by the Flanders Research Foundation (FWO) Research Grant No. FWO.OPR.2017.0014.01 on project No. G011617N.

References

Amenta, S., Marelli, M., & Crepaldi, D. (2015). Semantic consistency measures in priming paradigms. Paper presented at the 1st Quantitative Morphology Meeting, Belgrade, Serbia.

Amenta, S., Marelli, M., & Sulpizio, S. (2016). From sound to meaning: Phonology-to-semantics mapping in visual word recognition. *Psychonomic Bulletin & Review*, *24*, 887–893. doi:https://doi.org/10.3758/s13423-016-1152-0

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*, 439–461. doi:https://doi.org/10.3758/BF03214334

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models—Going beyond SVD. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 1–10). Washington, DC: IEEE Press. doi:https://doi.org/10.1109/FOCS.2012.49

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–481. doi:https://doi.org/10.1037/a0023851

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

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316. doi:https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. doi:https://doi.org/10.3758/BF03193014

Baroni, M., Bernardi, R., & Zamparelli, R. (2014a). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*, 241–346.

Baroni, M., Dinu, G., & Kruszewski, G. (2014b). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Association for Computational Linguistics (ACL)* (pp. 238–247). New York, NY: ACM Press.

Bates, E., Burani, C., d'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, *29*, 986–999. doi:https://doi.org/10.3758/BF03195761

Bernardi, R., Dinu, G., Marelli, M., & Baroni, M. (2013). A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 53–57). East Stroudsburg PA: ACL.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching word vectors with subword information*. Retrieved from arXiv: 1607.04606

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, *52*, 131–143.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. doi:https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991–997. doi:https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. doi:https://doi.org/10.3758/s13428-013-0403-5

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531–544.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256. doi:https://doi.org/10.1037/0033-295X.108.1.204

Davis, C. J. (1999). The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition (Doctoral dissertation, University of New South Wales, 1999). *Dissertation Abstracts International*, *62*(1-B), 594. Available from www.maccs.mq.edu.au/~colin.

Davis, C. J., & Bowers, J. S. (2004). What do letter migration errors reveal about letter position coding in visual word recognition?. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 923–941. doi:https://doi.org/10.1037/0096-1523.30.5.923

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*, 603–615.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*, 228–244. doi:https://doi.org/10.1016/0749-596X(90)90074-A

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*, 662–720. doi:https://doi.org/10.1037/0033-295X.111.3.662

Herdağdelen, A. (2013). Twitter *n*-gram corpus with demographic metadata. *Language Resources and Evaluation*, *47*, 1127–1147.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How facebook and twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, *41*, 976–995

Hino, Y., Miyamura, S., & Lupker, S. J. (2011). The nature of orthographic–phonological and orthographic–semantic relationships for Japanese kana and kanji words. *Behavior Research Methods*, *43*, 1110–1151.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*, 1744–1756.

Jared, D., Jouravlev, O., & Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 422–450.

Keith, J., Westbury, C., & Goldman, J. (2015). Performance impact of stop lists and morphological decomposition on word–word corpus-based semantic space models. *Behavior Research Methods*, *47*, 666–684.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304. doi:https://doi.org/10.3758/s13428-011-0118-4

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*, 1065–1081. doi:https://doi.org/10.1037/a0035669

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*, 211–240. doi:https://doi.org/10.1037/0033-295X.104.2.211

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*(Suppl. 4), 677–705.

Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 1517–1526). East Stroudsburg, PA: ACL.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods,*

Instruments, & Computers, *28*, 203–208. doi:https://doi.org/10.3758/BF03204766

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*, *97*, 57–78.

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

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), Proceedings of the 26th International Conference on Neural Information Processing Systems (Vol. 2, pp. 3111–3119). New York, NY: Curran Associates.

Milin, P., Divjak, D., & Baayen, R. H. (2017a). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1730–1751. doi:https://doi.org/10.1037/xlm0000410

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017b). Discrimination in lexical decision. *PLoS ONE*, *12*, e0171935. doi:https://doi.org/10.1371/journal.pone.0171935

Monaghan, P., & Christiansen, M. H. (2006). Why form-meaning mappings are not entirely arbitrary in language. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 1838–1843). Mahwah, NJ: Lawrence Erlbaum.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357. doi:https://doi.org/10.1037/0033-295X.113.2.327

Pecher, D. (2001). Perception is a two-way junction: Feedback semantics in word recognition. *Psychonomic Bulletin & Review*, *8*, 545–551. doi:https://doi.org/10.3758/BF03196190

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*, 161–167. doi:https://doi.org/10.3758/PBR.15.1.161

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*, 542–549.

Rodd, J. M. (2004). When do leotards get their spots? Semantic activation of lexical neighbors in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 434–439. doi:https://doi.org/10.3758/BF03196591

Samson, D., & Pillon, A. (2004). Orthographic neighborhood and concreteness effects in the lexical decision task. *Brain and Language*, *91*, 252–264.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Eds.), Natural language processing using very large corpora (pp. 13–25). Berlin, Germany: Springer.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384–394). East Stroudsburg, PA: Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176–1190. doi:https://doi.org/10.1080/17470218.2013.850521

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207. doi:https://doi.org/10.3758/s13428-012-0314-x

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. J. (2012). An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*, 72. doi:https://doi.org/10.3389/fnhum.2012.00072

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750. doi:https://doi.org/10.3758/s13423-011-0092-y

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979. doi:https://doi.org/10.3758/PBR.15.5.971